Institute for Cell and Molecular Biosciences

and the School of Computing Science

Comparative genomics for studying the proteomes of mucosal microorganisms

Sirintra Nakjang

Submitted for the degree of Doctor of Philosophy in the Institute for Cell and Molecular Biosciences, Newcastle University

December 2010

Acknowledgements

I would like to thank all those who helped and supported me throughout this thesis. In particular, I would like to thank my supervisors, Dr. Robert Hirt and Prof. Anil Wipat for their expert guidance. Prof. Anil Wipat provided valuable advice and support with the bioinformatics aspects of the work and Dr. Robert Hirt provided me with advice on the biological aspects of the thesis.

Many thanks also go to the Microbase developer, Dr. Keith Flanagan, whose advice in software engineering and technical support were invaluable for developing the high-throughput sequence analysis workflow. Thanks also for the help and advice I received from Dr. Jennifer Hallinan, Dr. Matthew Pocock, Dr. Phillip Lord and Dr. Daniel Swan. I would also like to thank the members of the National Centre for Text Mining at Manchester University for their implementation of the text-mining system to extract microorganism-habitat information.

I am also grateful to all the members of the Bioinformatics group at Newcastle University. In particular, I would like to thank Goksel Misirli, Katherine James, Jochen Weile and Morgan Taschuk who provided me with much inspiration. Special thanks go to the members of the writing group, whose weekly sessions picked up many typographical mistakes and helped me to improve the clarity of my writing.

I would like to extend my thanks to the Medical School at Newcastle University for providing financial support for my project.

Finally, I would like to thank my family for their love and for supporting me throughout the PhD course. Without their continued support, this thesis would not have been possible.

Declaration

I declare that this thesis, apart from the help recognised, is my own work. No part of this thesis has previously been submitted for a degree or any other qualification in this or another University. The work described in Sections 5.2.2, 5.3.2, 5.4.2 was conducted in collaboration with the National Centre for Text Mining (NaCTeM) and my contribution has been explicitly stated in these sections.

Sirintra Nakjang
February 2011

Abstract

A tremendous number of microorganisms are known to interact with their animal hosts. The outcome of the interactions between microbes and their animal hosts range from modulating the maintenance of homeostasis to the establishment of processes leading to pathogenesis. Of the numerous species known to inhabit humans, the great majority live on mucosal surfaces which are highly defended. Despite their importance in human health, little is known about the molecular and cellular basis of most host-microbe interactions across the tremendous diversity of mucosal-adapted microorganisms.

The ever-increasing availability of genome sequence data allows systematic comparative genomics studies to identify proteins with potential important molecular functions at the host-microbe interface. In this study, a genome-wide analysis was performed on 3,021,490 protein sequences derived from 867 complete microbial genome sequences across the three domains of cellular life. The ability of microbes to thrive successfully in a mucosal environment was examined in relation to functional genomics data from a range of publicly available databases. Particular emphasis was placed on the extracytoplasmic proteins of microorganisms that thrive on human mucosal surfaces. These proteins form the interface between the complex host-microbe and microbe-microbe interactions.

The large amounts of data involved, combined with the numerous analytical techniques that need to be performed makes the study intractable with conventional bioinformatics. The lack of habitat annotations for microorganisms further compounds the problem of identifying the microbial extracytoplasmic proteins playing important roles in the mucosal environments. In order to address these problems, a distributed high throughput computational workflow was developed, and a system for mining biomedical literature was trained to automatically identify microorganisms' habitats.

The workflow integrated existing bioinformatics tools to identify and characterise protein-targeting signals, cell surface-anchoring features, protein domains and protein families. This study successfully demonstrated a large-scale comparative genomics approach utilising a system called Microbase to harness Grid and Cloud computing technologies.

A number of conserved protein domains and families that are significantly associated with a specific set of mucosa-inhabiting microorganisms were identified. These conserved protein regions of which their functions were either characterised or unknown, were quite narrow in their coverage of taxa distribution, with only a few protein domains more widely distributed, suggesting that mucosal microorganisms evolved different solutions in their strategies and mechanisms for their survival in the host mucosal environments. Metabolic and biological processes common to many mucosal microorganisms included: carbohydrate and amino acid metabolisms, signal transduction, adhesion to host tissues or contents in mucosal environments (e.g. food remnants, mucins), and resistance to host defence mechanisms. Invasive or virulence factors were also identified in pathogenic strains. Several extracytoplasmic protein families were shared among prominent bacterial members of gut microbiota and microbial eukaryotes known to thrive in the same environment, suggesting that the ability of microbes to adapt to particular niches can be influenced by lateral gene transfer. A large number of conserved regions or protein families that potentially play important roles in the mucosa-microbe interactions were revealed by this study. Several of these candidates were proteins of unknown function. The identified candidates were subjected to more detailed computational analysis providing hypothesis for their function that will be tested experimentally in order to contribute to our understanding of the complex host-microbe interactions.

Among the candidates of unknown function, a novel M60-like domain was identified. The domain was deposited in the Pfam database with accession number PF13402. The M60-like domain is shared amongst a broad range of mucosal microorganisms as well as their vertebrate hosts. Bioinformatics analyses of the M60-like domain suggested a potential catalytic function of the conserved motif as gluzincins metalloproteases. Targeting signals were detected across microbial M60-like-containing proteins. Mucosa-related carbohydrate-binding modules (CBMs), CBM32 was also identified on several proteins containing M60-like domains encoded by known mucosal commensals and pathogens. The co-occurrence of the CBMs and M60-like domain, as well as annotated potential peptidase function unveiled a new functional context for the CBM, which is typically connected with carbohydrate processing enzymes but not proteases. The CBM domains linked with members of different protease families are likely to enable these proteases to bind to specific glycoproteins from host animals further highlighting the importance of proteases and CBMs (CBM32 and CBM5_12) in host-microbe interactions.

Contents

Abbreviations

CBM Carbohydrate-Binding Module

COGs Cluster of Orthologous Groups

DAGs Directed Acyclic Graphs

ECM Extracellular matrix

EnvO Environment Ontology

ExCyt Extracytoplasmic

FTP File Transfer Protocol

gbk GenBank-formatted

Gm+ Gram-positive bacteria

Gm- Gram-negative bacteria

GOLD Genomes OnLine Database

GPDB GenomePool database

GIT gastrointestinal tract

GPI Glycosylposphatidylinositol

GO Gene Ontology

HMM Hidden Markov Model

HMP Human Microbiome Project

IPR InterPro

KEGG Kyoto Encyclopedia of Genes and Genomes

LGT Lateral gene transfer

MIGS Minimal information about a Genome Sequence

NCBI National Center for Biotechnology Information

NN Neural Network

OBO Biomedical Ontologies

PF Pfam

PredExtDB Predicted extracytoplasmic protein database

RefSeq Reference Sequence

RT respiratory tract

S-layer surface layer

SIgA secreted immunoglobulin A

SLH Surface Layer Homology

SP signal peptide

SPI signal peptidase I

SPII signal peptidase II

SVM Support Vector Machine

TMH transmembrane helices

UGT urogenital tract

VMs Virtual machines

Chapter 1

Introduction

Proteins form numerous different types of structures and perform virtually all cellular functions essential for the survival of all cellular life forms. Hence, these polypeptides fundamentally determine the overall phenotypes of organisms. Phenotypes are expressed through regulation in time and space of protein expression and function. Knowing the subcellular localisation (space) of a protein is a fundamental information to determining its function [Gardy and Brinkman, 2006][Billion *et al.*, 2006] [Gardy, 2004]. For example, proteins located in cytoplasm can function as part of cytoskeleton or translation processes. Extracytoplasmic (transmembrane, surface-anchored and secreted) proteins of prokaryotes and microbial eukaryotes are known to play important roles in the interaction between the microbes and their biotic and abiotic niches. The main functions of extracytoplasmic proteins include nutrient acquisition, waste transport, degradation of extracellular compounds, binding to substrates and cell membrane, as well as cell communication [Lin *et al.*, 2002].

In the human microbiota, extracytoplasmic proteins play crucial roles in the interaction with the host including adhesion, invasion, signal transduction, evasion and modulating host immune responses [Pallen and Wren, 2007][Niemann, 2004]. Interestingly, human microbial communities have wide-ranging effects, from providing enormous benefits to our health to dramatic deterioration of our normal physiology [Turnbaugh *et al.*, 2007]. Human gastrointestinal microflora influence our normal physiology by providing structural, protective and metabolic functions [O'Hara and Shanahan, 2006]. For example, some gut commensals are known to help the renewal and differentiation of gut epithelial cells [Pull *et al.*, 2005]. Some are involved in several metabolic pathways such as vitamin synthesis and process polysaccharides that are not digestible by human [Martin *et al.*, 2007][Bäckhed *et al.*, 2005]. In contrast, some pathogenic strains such as enterohaemorrhagic *Escherichia coli* (EHEC) and *Entamoeba histolytica* cause hemorrhagic colitis and amoebiasis, respectively [Loftus *et al.*, 2005]

[Bielaszewska and Karch, 2005]. Experimentation and detailed studies are required to reveal important protein or genotypic features, driving the dynamic processes of the host-microbe interactions [Dethlefsen *et al.*, 2007]. However, genomics provides data to generate hypotheses and guide experimental work. The human microbiome project is underway with the aim of gaining a better understanding of these host-microbe interactions [Turnbaugh *et al.*, 2007].

1.1 Motivation for this project

The human body is a habitat for miriads of microorganisms and they reside in a wide variety of anatomical parts with, in particular, the skin and mucosal surfaces of the respiratory, gastrointestinal and urogenital tracts [Dethlefsen *et al.*, 2007][Costello *et al.*, 2009]. The human microflora form site-specific communities. The composition of microbial communities of these sites also varies considerably across individuals depending on their genetics [Benson *et al.*, 2010], age, diet, health status and medication history [Costello *et al.*, 2009][Kuczynski *et al.*, 2010]. Many mucosa-associated microorganisms are able to grow in habitats with diverse substrates such as the gastrointestinal tract. Digestion of various food products requires a broad range of enzymes, many of which are produced by the gut microbiota and are secreted or expressed on the cell surface [Sonnenburg *et al.*, 2005][Martens *et al.*, 2008]. The efficiency of food processing is also influenced by microbe-microbe interaction [Gill *et al.*, 2006]. In addition, the microbes must also deal with environmental pressures controlled by the host such as the secretion of gastric acidic juices and enzymes, dynamic flushing of mucosal secretions, and defence mechanisms of the host immune system [Nataro *et al.*, 2005].

The adaptation of microbes to particular habitats is a factor that shapes particular expressions of microorganisms' protein profiles [Ren and Paulsen, 2005]. As observed by many studies, variations of extracytoplasmic proteins are found across microorganisms in relation to the surrounding extracellular environment [Goh *et al.*, 2006][Rasko *et al.*, 2005][McMeechan *et al.*, 2005]. These adaptations are driven by several evolutionary mechanisms such as gene duplication, gene loss, horizontal gene transfer and genome rearrangement [Fraser-Liggett, 2005][Medini *et al.*, 2008]. The genetic basis of microorganism adaptation to a specific environment is typically poorly known. Particularly, the processes that enable a microorganism to thrive in mucosal environments, highly defended host compartments, are also not well known. Another open research question is what makes a particular microbe either colonise in a beneficial way, or aggressively invade and damage its hosts [Dethlefsen *et al.*, 2007][Turnbaugh *et al.*, 2007].

The increased availability of complete genome sequence data for numerous organisms across the three domains of cellular life allow more thorough comparative genomics study. Comparative genomics analyses can be carried out over a wide range of organisms with different traits to pinpoint sets of genotypes or protein complements underlying specific phenotypes [Ahmed, 2009]. Several comparative studies have successfully inferred groups of functional protein domains or families that are exclusively expressed on microbes in particular conditions or habitats [Goh *et al.*, 2006][Liu *et al.*, 2006]. Several of such studies were performed on a restricted set of taxa or taxonomic groups [Read *et al.*, 2003] [O'Sullivan *et al.*, 2009]. The work presented in this thesis has explored the potential associations between proteomes and the ability of microorganisms to thrive in a host mucosal environment to gain a better understanding of a molecular basis of host-microbe interactions. Of particular interest are essential fundamental aspects of the physical interactions of microbes with host mucosal surface barriers. To date, no published study has yet applied comparative genomics techniques to all domains of cellular life for this purpose.

This study focussed on mucosal microorganisms and functional elements important for their survival in highly-defended host mucosal environments. The extracytoplasmic proteomes of these microbes represent one of the main interest of the study as they are at the interface of these complex host-microbe interactions, as well as interactions between members of microbial communities. The study covers a diverse range of microbial relationships with hosts (mutualism, symbiotic and parasitism) as well as various mucosa-lined surfaces known as microbe-dwelling places such as the oral cavity, gastrointestinal, respiratory, and urogenital tracts.

1.2 Project aims

The main aim of this project was to gain a better understanding of the structural diversity and evolutionary forces that shape microbial extracytoplasmic proteomes across all forms of cellular life (bacterial, archaeal and eukaryotic) in order to identify important functional elements mediating host-microbe interactions.

The genome wide associations were performed between the extracytoplasmic protein complements of microorganisms and the environment in which they reside. In particular, emphasis has been placed on microorganisms that thrive on human mucosal surfaces. The results of the work will be used to inform future laboratory based studies and to gain a better understanding of microbe-mucosa interactions.

The working hypothesis is that analysis of a large number of annotated genome sequences throughout the three domains of life will make identification of the extracytoplasmic proteome more sensitive [Ahmed, 2009][Gardy, 2004]. Large numbers of taxa also make correlation analysis feasible to differentiate important proteins found in specific microbial communities within an ecological niche from other niches. These proteins are hypothesised to have presence or absence distribution patterns or modulation of gene family sizes that correlate with a mucosal lifestyle of microorganisms.

An example of an overrepresentation of a particular set of protein-coding gene families in water-living bacteria compared to non-waterborne free-living bacteria and host-associated bacteria was published by Audic *et al.* [Audic *et al.*, 2007]. In a more specific example for mucosal-associated organisms, the BspA-like proteins, sharing a specific type of leucine rich repeats (LRR), are distributed among several known mucosal microbes across all three domains of life [Hirt *et al.*, 2002][Noël *et al.*, 2010]. BspA is a surface protein functional characterised in *Tannerella forsythensis*, and some other taxa to have an important role in the colonisation of mucosal tissues, binding between microbes and inducing innate and adaptive host immune responses [Hirt *et al.*, 2007][Noël *et al.*, 2010].

1.3 Project objectives

- To identify through comparative analysis of protein contents, in particular, extracytoplasmic protein regions either commonly conserved across, or unique to mucosal microorganisms. Hypotheses should be generated regarding the involvement of these proteins in interactions with the host mucosa environment.

- To develop automated bioinformatics pipelines, using a Grid-based computational system to manipulate and analyse the large amount of data derived from genome sequencing projects. The pipeline was to integrate available genome databases, bioinformatics tools, algorithms and other related biological knowledge to serve the needs of the project. The Grid-based computational system would also be designed to allow updates to the computed data sets when new completed genome sequence data is released.

- To classify microorganisms by their environment of predilection in which they thrive by mining existing data from published literature. The results were to be used to allow comparative analyses designed to contrast the protein contents of mucosal microbes with those thriving in other environments.

1.4 Thesis structure

- Chapter 2 provides background and a literature review of previous work relating to this thesis.
- Chapter 3 describes the development of a high-throughput sequence analysis workflow used to perform the identification of microbial extracytoplasmic proteins, the recognition of known functional protein domains and the sequence homology search.
- Chapter 4 describes computational approaches for the identification of microbial extracytoplasmic proteins and presents the results of these analyses.
- Chapter 5 describes a procedure for the automated annotation of microorganisms' habitat information using a text-mining approach.
- Chapter 6 presents analysis of the *in silico* identification and characterisation of mucosa-associated proteins.
- Chapter 7 presents the bioinformatics analyses identifying a novel zinc-metalloprotease-like domain in host-associate microbes. These analyses also suggest a new functional context for carbohydrate binding modules.
- Chapter 8 discusses the overall approaches, results and potential future work.

Chapter 2

Background

2.1 Genomics

A genome sequence represents an entire nucleic acid-based genetic information of an organism and is comprised of three major components: non protein-coding genes, protein-coding genes and regulatory elements. Proteins are the products of genes. Microbial phenotypes are typically driven by the expression of protein-coding genes which mediate cellular mechanisms. These polypeptides express functional units essential for the survival of organisms in response to challenges in their environment. The combination and variation of genes and their products can cause noticeable differences among organism phenotypes.

One of the aims of genome analysis is to infer the phenotypic potential of organisms. More specifically, to gain a better understanding of the molecular functions of cells encoded by the genome sequences. Genome databases are currently growing at an exponential rate (see Figure 2.1) due to the advent of high-throughput sequencing technologies [Medini *et al.*, 2008]. For example, the May 2010 release of the UniProtKB/TrEMBL protein database added 161,141 new sequences. The total number of sequence entries made available was 10,706,472[1]. As a result, there is a tremendous amount of genome and proteome data available for annotation and study.

In addition to the considerable amount of information already available, a large collection of new genome sequence information from ongoing genome sequencing and metagenomics projects will become available in the near future, in particular, for microbes colonising mammalian hosts (e.g. Human Microbiome Project [Turnbaugh *et al.*, 2007]). This information will be invaluable for understanding the roles of different microbial communities on the host mucosal environments [Flint *et al.*, 2008].

[1] `http://www.ebi.ac.uk/uniprot/TrEMBLstats/`, accessed 5th May 2010

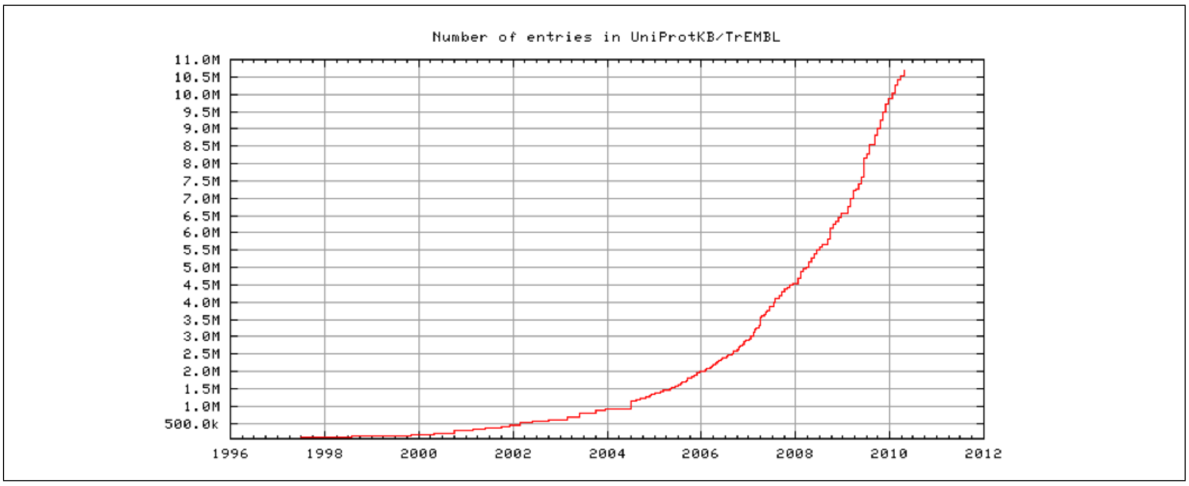


Figure 2.1: The growth of the UniProtKB/TrEMBL protein database. The number of well-annotated protein sequences deposited in the protein database grows dramatically at an exponential rate from 2004 until the present time. This figure was obtained from `http://www.ebi.ac.uk/uniprot/TrEMBLstats/` (accessed 5th May 2010).

Comparative genomics is one approach that can be used to investigate the associations between the genotypic features and the phenotypes of microorganisms from various taxonomic groups, as well as different ecological niches [Medini *et al.*, 2008].

2.2 Human microbiome and mucosal surfaces

A complete set of microorganisms that inhabit in a particular habitat are known as microbiota [Ley *et al.*, 2008]. The human microbiome comprises a collection of genes of the microbiota that live within the human body [Turnbaugh *et al.*, 2007]. The human body harbours a tremendous number of diverse microorganisms. These normal flora are found on the human surfaces that are exposed to the outside world, namely skin and mucosal surfaces. Microbial communities are also found on the areas where skin and mucosal epithelium are joined (mucocutaneous zone), such as the anus, nasal and ear cavities [Costello *et al.*, 2009].

2.2.1 Mucosa

Mucosa are protective layers coating several internal organs of vertebrates. The mucosa are covered or protected by a variety of secretions including mucus, immunoglobulins (mainly secreted immunoglobulin A (SIgA)), antimicrobial substances (e.g. lysozyme, lactoferrin, defensins) as well as normal flora

[Acheson and Luccioli, 2004]. This physical barrier is an interface to the external environment of several mammalian anatomical structures including the gastrointestinal tract (GIT), respiratory tract (RT), mammary gland and urogenital tract (UGT) [Nagler-Anderson, 2001] [Vélez *et al.*, 2007] (see Figure 2.2). The mucosal surfaces mediate exchanges vital for human homeostasis and reproduction including food processing and absorption (GIT), gas exchanges (RT), and waste removal (GIT, UGT). This barrier also a place where numerous and dynamic host-microbe interactions take place.

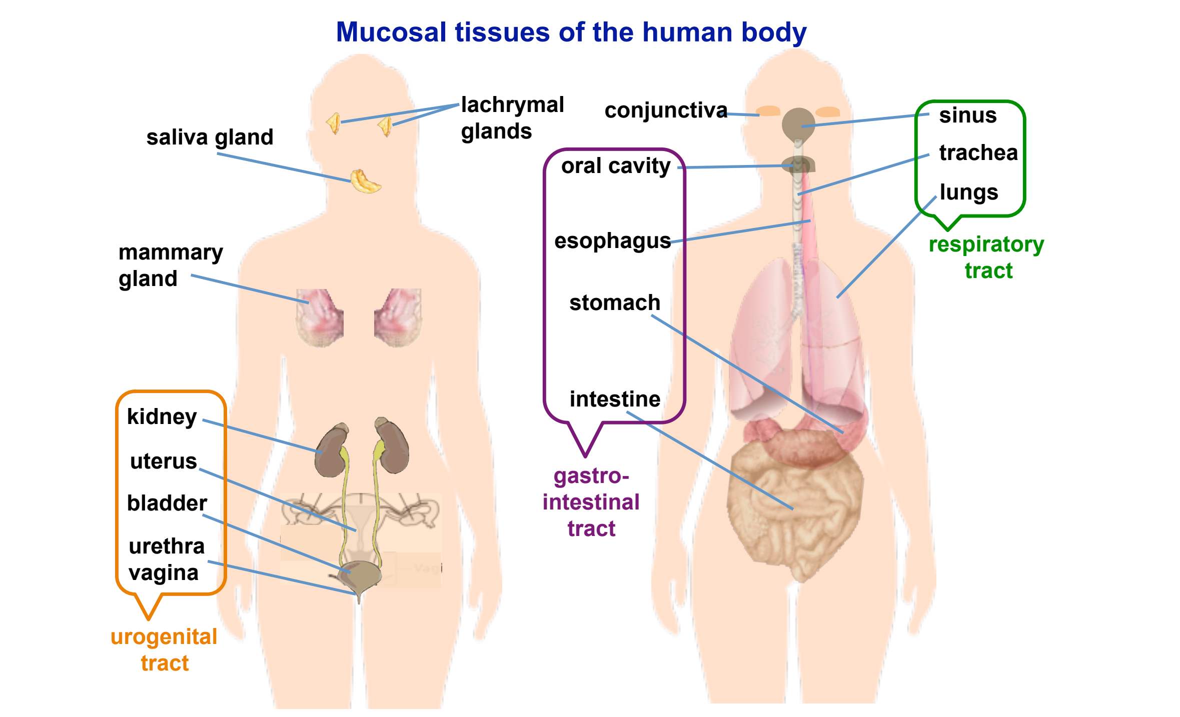


Figure 2.2: Mucosa-lined organs of the human body. The human mucosa-lined regions include epithelial membrane of urogenital, gastrointestinal and urinary tracts. Others are mammary gland and lachrymal gland and conjunctiva. The figure was adapted from Immunology 7th edition by Garland Science 2008. Each human organ part image were obtained from wikipedia's public domain: `http://en.wikipedia.org/wiki/File:Man_shadow_anatomy.png`.

One of the best understood mucosal environment is the human intestinal mucosa with several components including mucus layer, glycocalyx (carbohydrate-rich coating), epithelial cells, Extracellular matrix (ECM), lamina propria and muscularis mucosae (Figure 2.3) [Vélez *et al.*, 2007]. The ECM is composed of Type IV collagens laminins, fibronectin, tenascin-C, collagens and proteoglycans [Vélez *et al.*, 2007].

Mucus is mainly composed of mucins (high-molecular-weight glycoproteins) and minor components of lipids. Mucins consist of peptide backbones and O-glycosidically linked carbohydrate side chains. Studies of the composition of human gastric and bronchial mucins indicate the presence of the amino acids and carbohydrates. These amino acids include threonine, serine, proline, aspartic acid, leucine, glycine. The monosaccharides found to be enriches in mucins are fucose, man-

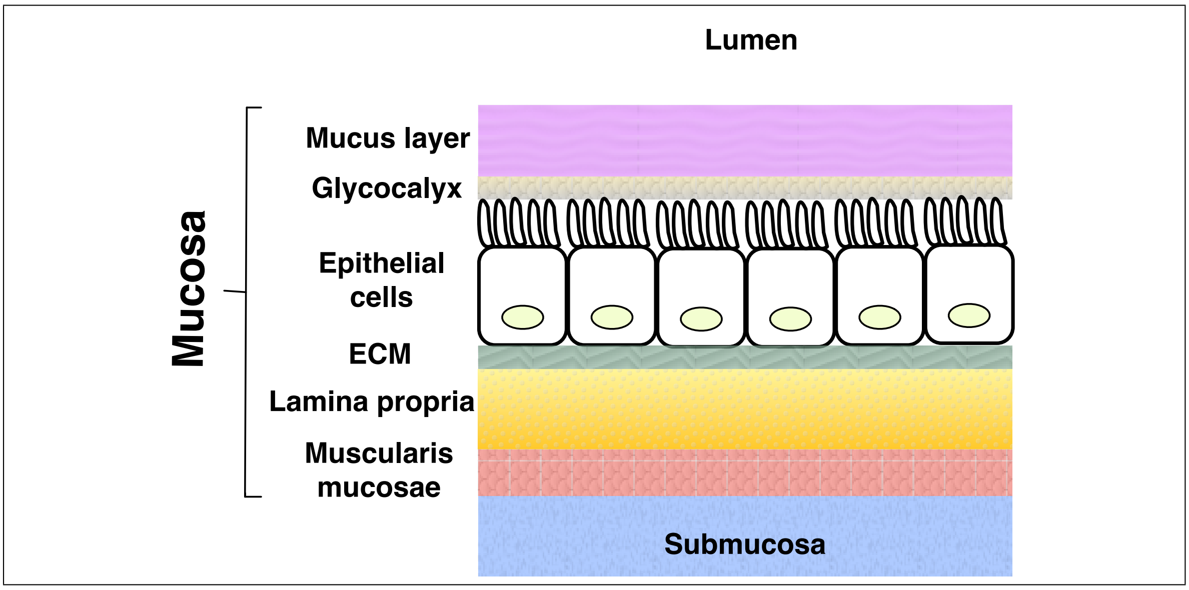


Figure 2.3: Overview of a mucosa architecture. The figure represents a simple model of the structure of human intestinal mucosa. Adapted from Velez M *et al* 2007 [Vélez *et al.*, 2007]. Mucosal surfaces are comprised of several sub-layers. This surface separate the external environment from the internal vertebrate organs. Multilayer of epithelial cell can be found on some other mucosa such as some part of the urinary tract [Nataro *et al.*, 2005].

nose, galactose, N-acetylgalactosamine, N-acetylglucosamine and sialic acid [Wagner *et al.*, 1998] [Bhattacharyya *et al.*, 1988].

2.2.2 The human microbiota and microbiome

The human microbiota is known to vary greatly over time [Costello *et al.*, 2009], depending on several factors of the host such as age, genetics, the status of the immune system, and lifestyle (e.g. diet) [Round and Mazmanian, 2009][Turnbaugh *et al.*, 2007][Acheson and Luccioli, 2004]. Different parts of the human anatomy harbour different microbial communities [Costello *et al.*, 2009]. The majority of the biomass of microbiota is located on mucosal surfaces. In particular, the large intestine is densely populated by microbial communities [O'Hara and Shanahan, 2006]. The human intestinal mucosa has an enormous surface area of roughly 400 m^2 [Acheson and Luccioli, 2004].The estimated number of cells of bacterial normal flora in a human body is in the region of 100 trillion. This number is about 10 times the total number of human cells and most of these microbial communities are in the human intestine [Bäckhed *et al.*, 2005][Gill *et al.*, 2006]. Interestingly, the human gut microbiome may contains more than 100 times the number of protein-coding genes in the human genome [Bäckhed *et al.*, 2005][Neish, 2009]. The human microbiome has provided us with important functional features that contribute to our health status [Neish, 2009][Blum and Schiffrin, 2003].

The human microbiota are known to contribute to our health and disease status through the complex host-microbe and microbe-microbe interactions (Figure 2.4 and 2.5). The human microbiota helps maintain our normal physiology. For example, bacterial colonisation of the gut promotes the development of our intestinal adaptive immune system [Round and Mazmanian, 2009]. An imbalance in composition of the mucosal microbiota can disrupt physiological processes and lead to disease. The study by Turnbaugh *et al.* (2006) indicated that changes in the two predominant bacterial divisions in the gut (Bacteriodetes and Firmicutes) is associated with obesity. An increase of the gut microbiota increases the capacity to acquire energy from the diet and may contribute to obesity [Turnbaugh *et al.*, 2006]. Several studies have shown that disturbances in the bacterial microbiota may underlie many disorders such as inflammatory bowel disease [Round and Mazmanian, 2009] [Qin *et al.*, 2010]. Recent molecular studies of the human gut microbiome reveal an immense diversity of the gut microbiota. Functions of prominent flora that are known to benefit the host body were reviewed by O'Hara A.M. and Shanahan F. [O'Hara and Shanahan, 2006]. However, many of the molecular mechanisms of host-microbe and microbe-microbe interactions, and their effects on our physiology remain unknown [Dethlefsen *et al.*, 2007][Ahmed *et al.*, 2007][Gill *et al.*, 2006]. A greater understanding of these complex interactions that underlie our homeostasis or pathophysiological status would provide valuable knowledge, enabling us to exploit the beneficial impacts and exert more control over the adverse effects. The understanding of how the human microbiota contribute to our health as well as diseases could eventually lead to the developments of probiotics as well as new therapeutic strategies [Sekirov *et al.*, 2010][Hattori and Taylor, 2009][Hooper and Gordon, 2001].

The Human Microbiome Project (HMP) was launched in 2007 with the aim of providing a better understanding of the role of human microbiota on human biology in terms of their contribution to health and disease [Turnbaugh *et al.*, 2007]. One aspect of the HMP project is to employ a comparative metagenomics approach to uncover the functional attributes of the microbiome. Microbiomes from various body surfaces (i.e. skin and mucosal surfaces) of several individuals are being collected, sequenced and analysed.

Based on 16S rRNA gene-sequence analysis, the human microbial communities have been found to be dominated by Firmicutes, Bacteriodetes, Proteobacteria, and Actinobacteria bacterial phyla [Dethlefsen *et al.*, 2007][Costello *et al.*, 2009]. Actinobacteria are found primarily on human skin, Firmicutes and Bacteriodetes are predominant on human mucosal surfaces (see Figure 2.6). Bacteria from the Firmicutes and Bacteriodetes phyla form the majority of the population found in human gut flora [Ley *et al.*, 2008]. However, other bacterial phyla are also found in the human GIT and UGT. These small proportions of bacterial phyla include Proteobacteria and Actinobacteria [Dethlefsen *et al.*, 2007]. Apart from bacterial species, the human microbiota also include archaea,

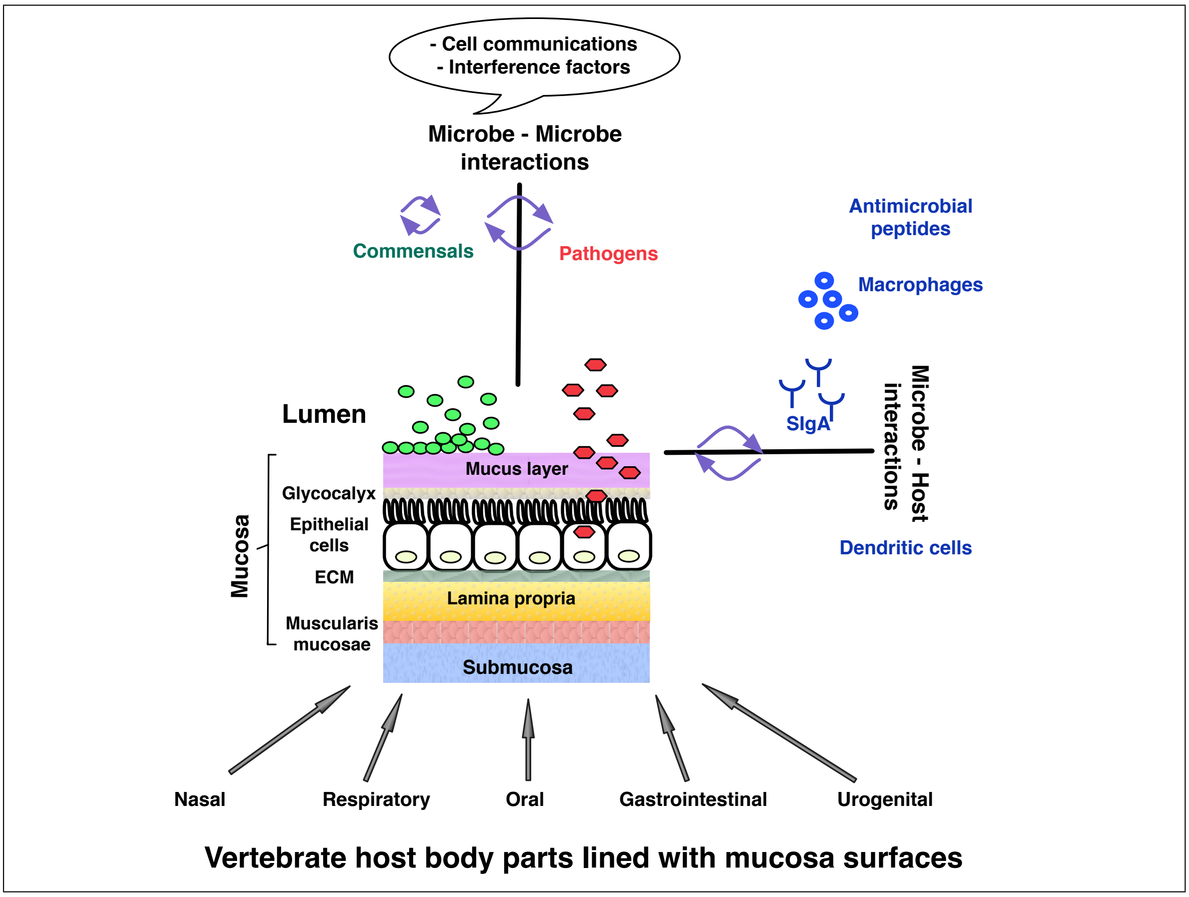


Figure 2.4: Overview of the mucosa-microbe interactions. A schematic representation of some of the complex interactions taking place at the mucosal surface include interactions between host cells or components, and microorganisms, as well as among members of the microbiota. Host elements that interact with microbes are mainly part of the host defence mechanisms such as macrophages, antimicrobial peptides, SIgA and mucins. Cell communications and metabolic co-operation are important processes for the survival of local microbial communities. Commensals also play important roles in defending their communities from invading microbes including pathogenic strains.

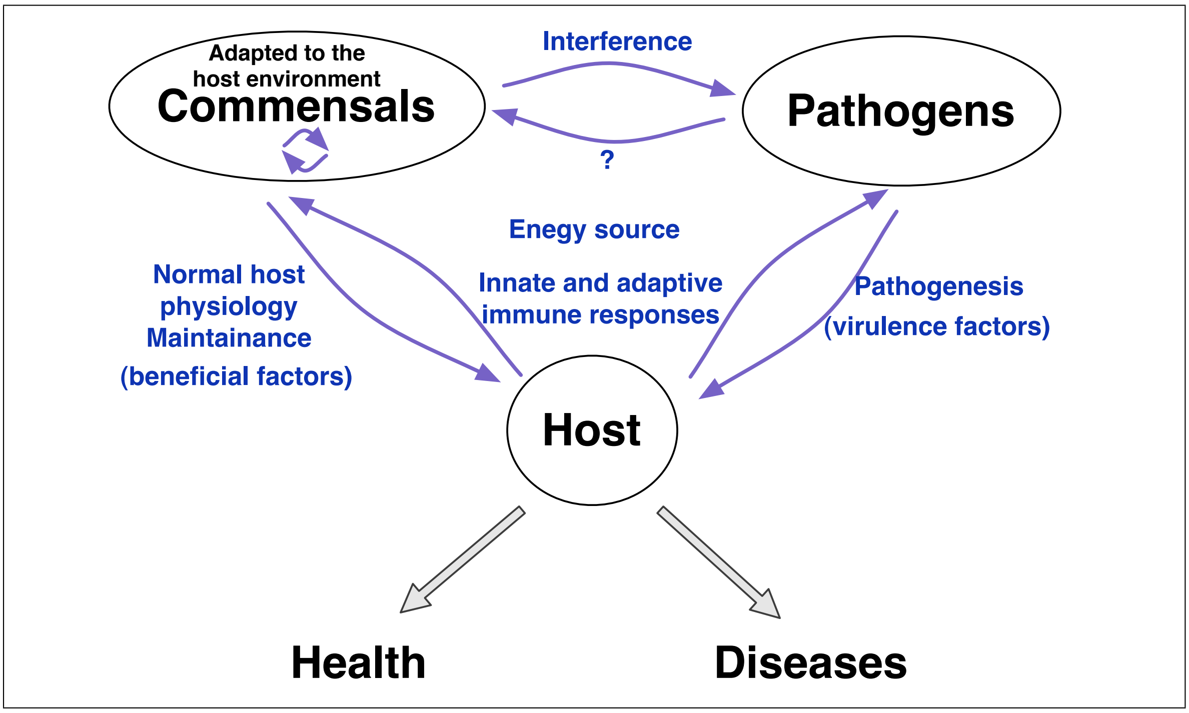
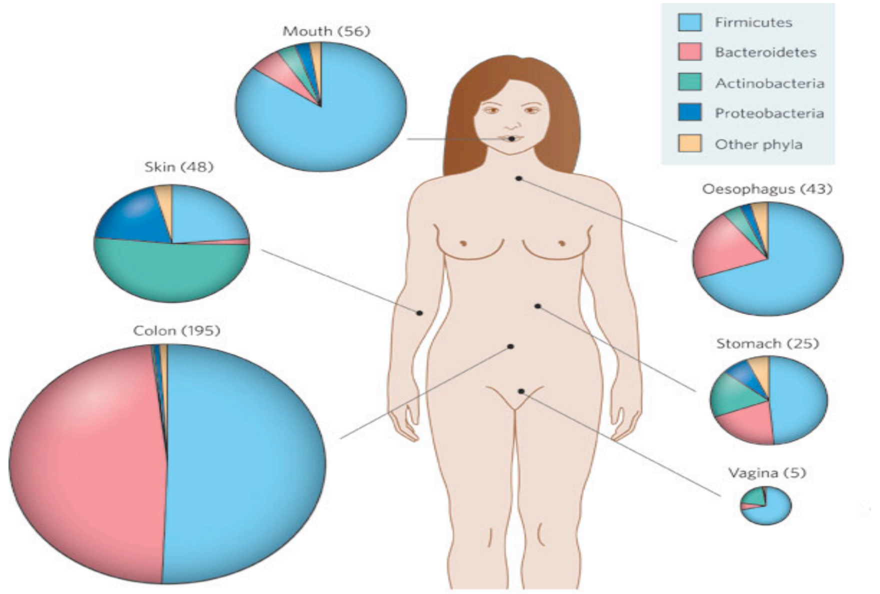


Figure 2.5: Molecular interactions of host, commensal and pathogenic microbes. The results of host-microbe interaction affect homeostasis status of the host body. Commensals help maintaining host normal physiology by, for instance, providing nutrients and boosting host defensive mechanisms. In return, the host environment provides a good source of energy to these local microbial communities. Moreover, these normal flora also act as secondary shields protecting their host from pathogens. The balance of these interactions is required in order to maintain a healthy stage of the host.

microbial eukaryotes and viruses [Reyes *et al.*, 2010]. These human microbial communities must adapt to live in a highly defended host environment. Studies of the large intestinal bacterial communities suggested that type and amount of host dietary intake have a major influence on the composition and metabolisms of various normal flora populations within the colon [Duncan *et al.*, 2007]. Dietary intake can shape gut flora communities and can be explained by the substrate preferences and competitive abilities among the gut microbial members [Flint *et al.*, 2008]. The variety of substrates originating from the host diet or mucus glycans influences the diversity of the ecological niches that can be exploited by the gut communities [Sonnenburg *et al.*, 2005].

2.2.3 Mucosal immunity

Vertebrates have evolved several defence mechanisms to protect themselves against foreign bodies, including microorganisms. These mechanisms are known as innate and adaptive immunities. The innate immunity is a non-specific defence system consisting of three main aspects: mechanical, chemical and cellular [Murphy *et al.*, 2007][Nataro *et al.*, 2005]. The mechanical aspect includes anatomical barriers (e.g., skin and mucosa) and movement of body parts (e.g. cilia, intestine) or



From An ecological and evolutionary perspective on human microbe mutualism and disease; Les Dethlefsen, Margaret McFall-Ngai & David A. Relman*Nature* **449**, 811-818(18 October 2007)doi:10.1038/nature06245

Figure 2.6: Site-specific distributions of bacterial phyla in healthy humans. The size of the chart represents the average number of distinct microorganism species per individual based on 16S rRNA gene-sequence survey. The average number per habitual site is shown in parenthesis. 3-11 healthy individuals per habitat were studied. The coloured wedges indicate the proportion of bacterial species regarding different phyla. The figure was derived from [Dethlefsen *et al.*, 2007].

contents (mucus, fluids). The chemical aspect of the defence arises from antimicrobial proteins, enzymes and sensor systems that recognise patterns of molecules. The cellular component of the innate immune system is composed of epithelial cells, phagocytes and normal flora [Nataro *et al.*, 2005].

The most well known positive effects of human normal flora are the intestinal commensals. The intestinal microflora provide a number of benefits to the human body. O'hara and Shanahan (2006) group these beneficial effects into three categories: protective, structural and metabolic functions [O'Hara and Shanahan, 2006] (Table 2.2).

Table 2.1: Mucosal innate immunity. Examples of different types of innate immunity found on human mucosa. The mucosal immunity is classified into three types, mechanical, chemical and cellular. (Adapted from [Murphy *et al.*, 2007])

	Gut	**Lungs**	**Eyes/nose/oral**	**Vagina**
Mechanical	Epithelial cells jointed by tight junctions			
	Peritalsis	cilia movement	Tears/Nasal cilia, Saliva flow	Urine flow
Chemical	Low pH, Digestive enzymes (pepsin, pancreatic enzymes, bile acids)		enzymes in tears and saliva (e.g. lysozyme)	Low pH
	Antimicrobial peptides			
Cellular	Normal flora (not for eyes), Host immune cells e.g., macrophages			

Table 2.2: Functions of human intestinal microflora beneficial to host body. This table provides examples of the known beneficial functions of intestinal flora. These functions are classified into three groups: protective, structural and metabolic functions. (Adapted from [O'Hara and Shanahan, 2006])

Functions	**Positive effects**
Protective functions	pathogen displacement, nutrient competition, receptor competition, production of anti-microbial factors e.g., bacteriocins, lactic acids
Structural functions	barrier fortification, induction of IgA, apical tightening of tight junctions, immune system development
Metabolic functions	Control IEC differentiation and proliferation, metabolise dietary carcinogens, synthesise vitamins e.g., biotin, folate, Ferment non-digestible dietary residue and endogenous epithelial-derived mucus, ion absorption, salvage of energy

The adaptive immunity responds to specific foreign materials by learning and remembering the most effective response to that particular material. This type of immune response is specific to each type of antigen [Murphy *et al.*, 2007].

2.3 Microbial cell surfaces and protein translocations

The project described in this thesis performed an analysis of the extracytoplasmic proteomes from a wide range of microorganisms including Archaea, Bacteria and microbial eukaryotes. Each group of these organisms show differences in their cell surface structures and protein translocation mechanisms. In this section, the following aspects are described:

- the major differences in the cell surface structure of Archaea, Bacteria and microbial eukaryotes;
- the important functions of the microbial extracytoplasmic proteins;
- a summary of the known protein secretion systems.

2.3.1 Diversity of cell surface structure

The cell surface is a selectively permeable barrier and the physical boundary of a cell. The structure of the cell surface is different across the diversity of all forms of cellular life. Prokaryotic and eukaryotic organisms have very distinctive cell surface features. The cell surface comprises either one or two membranes composed of lipids. There may also be a cell wall as the outermost layer. The chemical composition and topology of each part varies across taxonomic groups. Membrane lipids are mainly composed of carbon chains linked to glycerols. Membranes of both bacteria and eukaryotes, contain straight carbon chains are attached to glycerol molecules by ester linkages [Alberts *et al.*, 2007]. However, membrane lipids of archaea contain branched carbon chains that are bound to glycerol by ether linkages [Golyshina and Timmis, 2005]. Cell walls of bacteria contain peptidoglycan, whereas eukaryotic and archaeal cell walls lack peptidoglycan. Eukaryotic cell walls contain carbohydrates, which differentiate them from prokaryotic cell walls, for instance, the cellulose cell wall of plants [Lerouxel *et al.*, 2006] and the chitin cell wall of fungi [de Nobel *et al.*, 2000]. However, cell surfaces of some organisms such as *Ferroplasma* (archaea) [Golyshina and Timmis, 2005], *Mycoplasma* (bacteria) [Desvaux *et al.*, 2006], many microbial eukaryotes and animal cells do not have a cell wall. On the other hand, a specific surface layer such as a glycocalyx or surface layer (S-layer) may coat some bacterial and archaeal cells as well as animal cells [Frey, 1996]. The surface of Gram-negative bacteria consists of two layers of lipid bilayer membranes, an inner and an outer membrane, and a thin peptidoglycan layer in the periplasmic space [Alberts *et al.*, 2007]. In contrast, Gram-positive bacteria have only one plasma membrane surrounded by a thick peptidoglycan cell wall, usually containing teichoic acid [Desvaux *et al.*, 2006]. The diversity of cell surfaces (Figure 2.7) parallels diversity in protein secretion pathways and also determines how surface proteins are anchored to the cell surface.

2.3.2 Extracytoplasmic proteins

Expression of protein-coding gene sequences through regulation in time and space fundamentally convey the overall phenotypes of an organism. Proteins are synthesised in the cytoplasm and then either remain in this compartment, or are targeted to the cell surface or secreted to the external environment. Knowledge of the subcellular localisation (space) of a protein provides a clue to determine its biological function. For example, cytoplasmic proteins are more likely to be part of a cytoskeleton or translation process, whereas several extracytoplasmic proteins are known to mediate interaction between an organism and its surrounding environment. In this thesis, extracytoplasmic proteins include extracellular proteins (secreted or surface-anchoring proteins) as well as proteins exposed to the non-cytoplasmic compartment such as transmembrane proteins and outer membrane proteins (see Figure 2.8).

The main general functions of extracytoplasmic proteins of microorganisms include nutrient acquisition, waste transport, signal transduction, membrane and protein binding, as well as degradation of extracellular compounds. In terms of symbiosis and pathogenesis, these proteins are important for adhesion and biofilm formation, signal transduction, pathogen interference, invasion, and evasion. The microbial secreted proteins can act as enzymes involving the both microbes and host metabolic processes. Particularly for mucosal-thriving microorganisms, extracytoplasmic proteins are crucial, for instance, for degrading or binding to mucus, ECM proteins, epithelial cells and modulating the host innate and adaptive immune systems [Hirt *et al.*, 2002][Hirt *et al.*, 2007]. Several studies have showed that virulence factors were presented in the secretomes of pathogenic strains [Trost *et al.*, 2005]. For example, for invasion (e.g. Internalin A and B of *Listerria monocytogenes* [Trost *et al.*, 2005][Marino *et al.*, 2002]), for adhesion (e.g., TCP pili of *Vibrio cholerae* [Herrington *et al.*, 1988][Peterson and Mekalanos, 1988]), internalisation (e.g., invasin of *Yersinia pseudotuberculosis* [Isberg and Falkow, 1985][Isberg *et al.*, 1987]) and for defence against the host immune system (e.g., exotoxins of *Staphylococcus aureus* [Dinges *et al.*, 2000]). Some commensal strains such as *S. epidermidis*'s extracellular serine proteases (Esp) have been shown to inhibit biofilm formation and nasal colonisation by the pathogenic *S. aureus* [Iwase *et al.*, 2010].

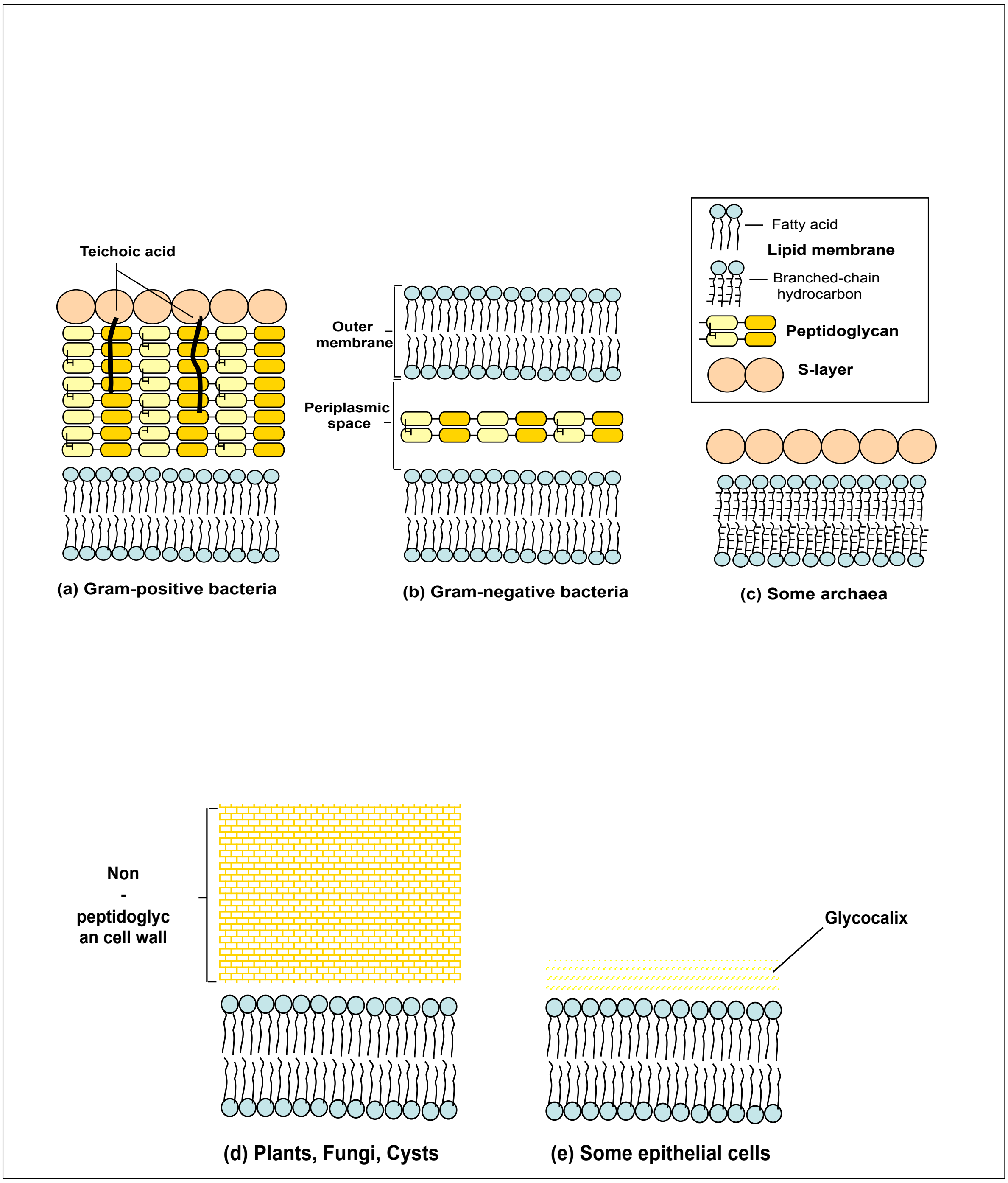


Figure 2.7: The cell surface and membrane organisations of various prokaryotic and eukaryotic cells. Surface proteins can be anchored on various components of the cell surface, including peptidoglycans, proteins of the S-layer and the cell plasma membrane. The glycocalix of animal cells, for example, is made of glycoproteins and glycolipids. Some microbial eukaryotes, such as Fungi, may have rigid cell wall made of chitin. Some eukaryotes have life cycle stages made of cysts or spores which have rigid protective coats which may consists of a mix of proteins, chitins, or other polysaccharides).

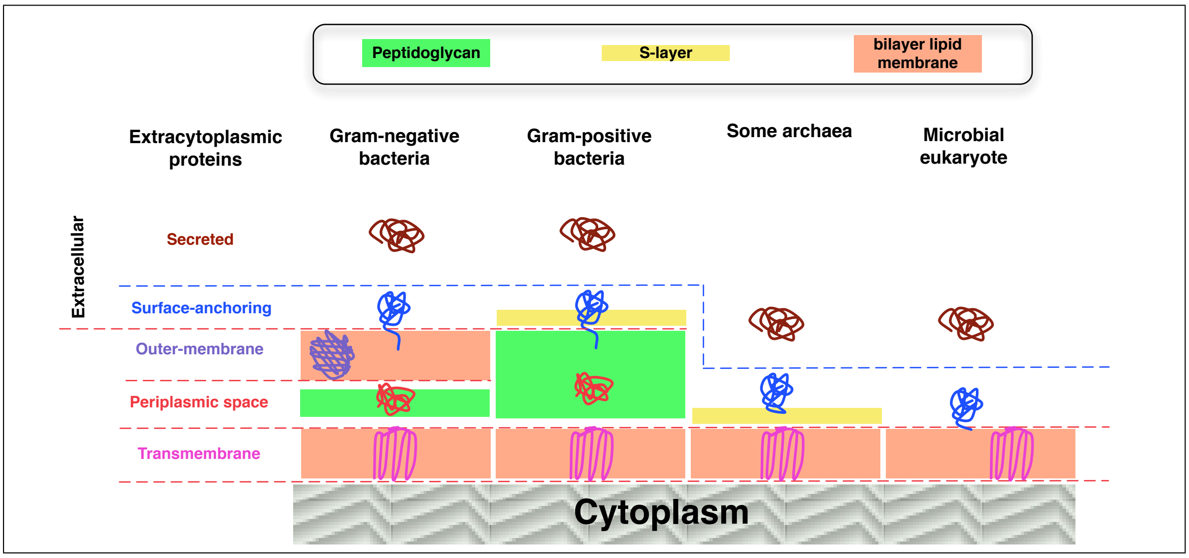


Figure 2.8: Subcellular locations of extracytoplasmic proteins of various microorganism cell surface structures. The cell surface structures of bacteria, archaea and microbial eukaryotes are shown. for each structure, possible locations of non-cytoplasmic proteins are indicated and terms used to refer to these proteins are provided. Terms mentioned in this diagram were used throughout this thesis.

2.3.3 Protein translocation mechanisms

Newly, synthesised proteins can be exported or translocated to various sites through one of several transport pathways, depending on structure and chemical composition of the cell surface which differ across the diversity of cellular life. Until now, several universal secretory systems have been well characterised across all three kingdoms of life (see Table 2.3). Transmembrane proteins are anchored to membranes by transmembrane helices (TMH) or as Beta-barrel proteins. However, some proteins have anchoring features that allow them to attach to the surface layer, resulting in the bulk of the protein being exposed extracellularly. Examples of such cell surface anchoring motifs are LPXTG (Gram-positive bacteria), S-layer motifs (Gram-positive bacteria) and Glycosylposphatidylinositol (GPI)-anchors (Eukaryotes) [Pallen *et al.*, 2003][Billion *et al.*, 2006][Bütikofer *et al.*, 2001].

The major secretion systems (described in Table 2.3) imply the presence of some recognisable features on the sequences themselves, allowing proteins to be targeted to a specific transport system. These features are generally defined as targeting-signals [Alberts *et al.*, 2007]. The trends of these amino acid residue compositions enable protein subcellular localisation predictions to be made by using computational methods to identify given sequence features. Moreover, depending on the cell surface structure, surface proteins can be anchored in different ways. Some conserved functional domains from well-characterised extracellular proteins can also be used to infer protein location. Examples are listed in Table 2.4.

However, not all virulent proteins are secreted through a classical secretory pathway [Bendtsen *et al.*, 2005b].

Table 2.3: Summary of major protein secretion systems and their distribution among the three domains of microbial cellular life. Sec = general or classical secretion pathway, Tat = twin-arginine translocation pathway, T1SS = type 1 secretion system, T2SS = type 2 secretion system, T3SS = type 3 secretion system, T4SS = type 4 secretion system, T5SS = type 5 secretion system, T6SS = type 6 secretion system

Secretory system	**Bacteria**	**Archaea**	**Eukaryote**	**Short description**
Sec	Yes	Yes	Yes	Transports proteins from the cytoplasmic space across the cell membrane into the extracellular compartment or topologically equivalent compartment [Pohlschroder *et al.*, 2005b][Pohlschröder *et al.*, 2005a].
Tat	Yes	Yes	Yes (only plastids)	
T1SS	Yes	No	No	Transports proteins from the cytoplasm directly to extracellular space using ATP-binding cassette (ABC) transporters [Lee and Schneewind, 2001].
T2SS	Yes	No	No	Transports proteins across or into the cell membrane either via the Sec-dependant pathway or the Tat pathway [Pallen *et al.*, 2003].
T3SS	Yes	No	No	Injects proteins directly into the cytoplasmic space of other cells via a pilus-like structure. A set of genes encoding protein subunits involved in the machinery of these systems is commonly transferred horizontally between pathogenic bacteria (known as Pathogenicity islands [Deng *et al.*, 2004]). However, the type IV system is not widely distributed among Gram-negative bacteria [Pallen *et al.*, 2003].
T4SS	Yes	No	No	
T5SS	Yes	No	No	The largest protein secretion system in Gram-negative bacteria. Also called autotransporters since secreted proteins are forced by an intrinsic activity of the substrate proteins. Proteins secreted through this system possess N-terminal signal peptides, targeting them to the Sec pathway before being translocated outside the cell [Pallen *et al.*, 2003].
T6SS	Yes	No	No	An important protein transportation system of virulent factors of Gram-negative pathogenic bacteria [Bingle *et al.*, 2008].

Table 2.4: Summary of major protein features that indicate cell surface or secreted proteins. Example are given for each feature. Available accession numbers are also listed.

Secretory signals	**Short description**
N-terminal targeting signals [Pallen *et al.*, 2003] [Pohlschröder *et al.*, 2005a] [Juncker *et al.*, 2003]	
Sec-signal sequences	A short sequence of hydrophobic amino acids targets nescent (unfolded) proteins to Sec secretory pathway. Also known as the classical N-terminal signal sequence distributed among all three domains of life
Tat-signal sequence	Shuttles folded proteins to the Tat secretory pathway. Consensus amino acid sequence is [S/T]RRxFLK.
Lipoprotein signal sequence (LPP)	Allows proteins to be exported and anchored covalently to the cell membrane [Sutcliffe and Russell, 1995] [Juncker *et al.*, 2003].
YSIRK (Pfam:PF04650)	YSIRK is characterised as a motif of Staphylococal protein A. A motif resembling [YF]SIRKxxxGxxS[VIA] appears at the start of the transmembrane domain. The motif facilitates a processing of signal peptide in the protein precursors and protein secretion but is not necessarily required [Bae and Schneewind, 2003].
Anchoring structure [Desvaux *et al.*, 2006] [Cabanes *et al.*, 2002]	
LPXTG motif (Pfam:PF00746)	Anchors proteins covalently to peptidoglycan of Gram-positive bacterial cell wall [Pallen *et al.*, 2003].
GW module (superfamily:22279)	The motif dipeptide Gly-Trp allows proteins to be anchored the bacterial cell surface via the interaction between the conserved module and lipoteichoic acids in Gram-positive bacterial walls. The domain, mediated by the carboxy-terminus, is non-covalently attached to the peptidoglycan or cytoplasmic membrane [Marino *et al.*, 2002].
Alpha-helical transmembrane domain	Usually contains a 15-30 hydrophobic amino acid residues long region and followed by positively charged residues. Presented in most transmembrane proteins from all domains of life [Krogh *et al.*, 2001].
Beta-barrel motif	Beta-berrel is a known structural motif for several protein spanning outer membrane of Gram-negative bacteria. The motif can also be found in the outer membranes of mitochondria and chloroplasts. Known beta-barrel structures contain between 8 and 22 beta strands [Wimley, 2002].
LySM (Pfam:PF01476)	LySin Motif (LySM) domain allows proteins to bind to peptidoglycan.
CWBD1 (Pfam:PF01473)	Non-covalent cell wall binding domain binds to the cell wall of Clostridium and Lactobacillus.
CWBD2	Non-covalent cell wall binding domain found in *B. subtilis* and *C. difficile*
S-layer motif (Pfam:PF00395)	Form strong binding non-covalent bond onto the cell surface peptidoglycan. Found in some archaea and bacteria [Billion *et al.*, 2006] [Desvaux *et al.*, 2006].
PKD-like domain	Found as an anchor in Archaeal surface proteins as well as eukaryotic membrane proteins.
GPI-anchored protein	GPI-anchored protein forms covalent bonds allowing a protein to attach to the outer part of cell membrane. Found in most eukaryotes [Omaetxebarria *et al.*, 2007]. Also found as an important key in host-microbe interaction as induce several host immune cells [Bütikofer *et al.*, 2001].
Examples of functional domain characteristics of some on surface and extracellular proteins	
Leucine-rich repeat (LRR)	Some LRR domains are expressed on the surface of prokaryotic [Cabanes *et al.*, 2002] and eukaryotic cells [Hirt *et al.*, 2002]. Four subfamilies of LRR (out of seven) are known to be characteristic on extracellular proteins [Kobe and Kajava, 2001].
NLPC/P60 domain (Pfam:PF00877)	Found in several prokaryotic surface and secreted proteins [Cabanes *et al.*, 2002].
Kringle domain	Presented in both kinetoplastid and apicomplexan extracellular proteins [Templeton, 2007].

For example, ESAT-6 from *M. tuberculosis* is secreted without the presence of typical signal sequences [Bendtsen *et al.*, 2005b][Bendtsen *et al.*, 2005a]. Since eukaryotic cells contain several subcellular organelles with each having their own specific membranes, translocation of proteins across those membranes requires appropriate protein sorting, targeting and retention signals. Many eukaryotic proteins with signal peptides are retained in membrane sealed organelles within the cell. A correlation between the proportion of different secreted proteins and the similarity of the environments in which the bacteria dwell has been observed [Bendtsen *et al.*, 2005a]. Lateral gene transfer (LGT) of gene-encoding virulent proteins occurs among pathogenic bacteria sharing the same hosts [Deng *et al.*, 2004] [Hsiao, 2003] [Garcia-Vallve *et al.*, 2003]. Data supporting LGT from prokaryotic to and among parasitic protozoa have also been published [Hirt *et al.*, 2007] [Carlton *et al.*, 2007] [Andersson, 2009].

2.4 Comparative genomics

Comparative genomics is a study of the association between genetic elements and organisms' phenotypes by comparing the genome information across species or strains to identify both conserved and divergent elements that express particular characteristic features of organisms. Comparative genomics studies can provide insights into the understanding of features that are essential for a species to survive in a habitat, particularly its indigenous habitat [Lee *et al.*, 2008]. This section contains a review of several comparative genomics works that were conducted using different data sources and statistical techniques.

2.4.1 Microbial Genotype-phenotype association analysis

The availability of completed microbial genomes and their phenotypic annotation provide the opportunity to understand the genetic basis of a trait by revealing the pattern of variation in gene or protein distribution [Jim, 2003]. The association between an organism's genome data and its phenotypic trait provides clues for understanding elementary biological mechanisms. The co-occurrence of genes and phenotypes were examined mostly based on combining phylogenetic profiles and phenotype profiles, followed by statistical approaches [Jim, 2003][Goh *et al.*, 2006][Slonim *et al.*, 2006]. Several statistical models have been employed to evaluate genotype-phenotype association and to pinpoint a significantly strong association or correlation between the genotypic information and traits of interest. Previous studies have been reviewed and summarised as follows.

Jim *et al.* (2003) [Jim, 2003] shows that reliable associations of genes and simple specific traits such as the presence of flagella or pili can be achieved by computing propensity scores, which allow less conserved proteins among organisms sharing the observed phenotype to be discovered. However, this approach seems to be limited by the frequency and specificity of the phenotype and whether the phenotype can emerge from more than one mechanism [Jim, 2003]. Comprehensive statistical methods based on Pearson's correlation coefficient and the hypergeometric distribution have proved successful in identifying pairwise associations between phenotypic laboratory results and the functional annotations of bacterial genomics contents (such as Cluster of Orthologous Groups (COGs), Gene Ontology (GO) annotations, Pfam (PF) entries, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways) [Goh *et al.*, 2006][Liu *et al.*, 2006]. Slonim *et al.* (2006) introduced a computational information-theoretic framework to extract clusters of genes that have a significant pairwise correlation with an observed trait. In this study, gene-phenotype associations were estimated using a statistical method called mutual information [Slonim *et al.*, 2006]. Their studies successfully demonstrated a systemic method for gathering phenotypic characteristics from the literature across a diverse set of species in order to associate them with genotypes. The approach revealed many novel trait-gene relationships, particularly for infectious disease-related genes [Korbel *et al.*, 2005].

Liu *et al.* (2006) [Liu *et al.*, 2006] correlates 63 microbial phenotypes among 59 prokaryotic species. This study applied a hypergeometric distribution function to find the probabilities of correlations between microbial phenotypes and functional genomics data. The genomics data included Pfam protein domains, COGs, KEGG pathways and GO.

Several comparative genomics studies were conducted with the aim to reveal genotypic signatures that differentiate human gut microbiome from the non-gut strains.

Lee *et al.* (2008) [Lee *et al.*, 2008] compared two *Bifidobacterium longum*'s genomes: one intestinal isolated and one cultured in the laboratory. The study identified regions of large deletion or gene loss in the cultured strain. The deleted regions were experimentally illustrated to be susceptible to deletion while growing outside the gut. These targeted gene sets are found uniquely in the gut-isolated Bifidobacteria which is involved in diverse traits pertinent to the human intestinal environment, specifically oligosaccharide and polyol utilization, arsenic resistance and lantibiotic production.

Sullivan *et al.* (2009) [O'Sullivan *et al.*, 2009] performed a comparative genomics study on 11 lactic acid bacterial genomes: 5 isolated from the human gut environment; 3 from diary products and 3 found to be present in multiple niches. This study identifies a unique gene set common to the gut-isolated species as well as the non-gut associated species by manual pairwise comparisons and sequence homology searches using BLAST. The authors proposed a barcode of 9 genes for the indication of the organism's ability to occupy a specific niche: 3 gut-specific genes and 6 diary-specific genes. The lactic acid bacterial gut-specific genes are involved in bile salt hydrolysis and sugar metabolism, while the diary-specific genes are part of proteolytic system and restriction/modification system.

Recently, the number of completed microbial genomes deposited in the GenBank database passed the 1,000 mark, and more than 1,000 others are currently in progress [Sayers *et al.*, 2010]. As more genome data becomes available, studying the loss and gain of entire groups of genes specific to a given phenotypic description becomes more feasible, allowing easier and more fine-grained analysis of genotype-phenotype correlations. However, an automated-systematic framework for high-throughput data analysis is required to manipulate the vast amounts of available data. Novel hypotheses can then be generated *in silico* and tested experimentally thereafter. An overview of high-throughput data analysis framework covered in this thesis is described in Section 2.8.4.

2.5 Microorganism-habitat information resources

Microorganisms play a significant role in symbiotic relationships ranging from commensalism, mutualism to parasitism [Steinert *et al.*, 2000]. To gain greater insights into the mechanisms involved in host-microbe interactions, it is essential to be able to contrast the genotypic features of microorganisms from various sources where microbes live, including both host-associated and non-host environmental niches.

Correlating genome content with microorganisms' ecological niches is of central importance to an understanding of the relationship between genotypes and phenotypes. However, one of the important limiting factors is assigning functional relevance to genome sequence data [Hirschman *et al.*, 2008] [Pallen and Wren, 2007]. The lack of resources providing information for genome sequence, such as isolation sources, was recently specifically recognised by the Minimal information about a Genome Sequence (MIGS) specifications [Field *et al.*, 2008]. In addition, several researchers have discussed or applied initial approaches to address this issue [Hirschman *et al.*, 2008][von Mering *et al.*, 2007].

To date, there is no complete computationally-accessible, structured data source for information regarding the habitat or isolation source of microorganisms whose genome sequence data is available. The National Center for Biotechnology Information (NCBI)[2] and the Genomes OnLine Database (GOLD)[3] databases are two most well-known public resources where information describing taxa can be obtained in the form of flat files. The information they provide includes isolation sources, habitat, organisms' morphology, motility, oxygen respiratory, endospore formation etc. However, this textual information is not always accessible for every microorganism whose genome data are available. More formally structured and detailed habitat information is important for comparative genomics studies and hence is required. A proposed solution for obtaining habitat information for the increasing number of organism-habitat pairs in an automated fashion to fulfil the need of a large-scale comparative genomics study is discussed in Chapter 5.

[2] `ftp://ftp.ncbi.nih.gov/genomes/genomeprj`, accessed 10th May 2010

[3] `http://www.genomesonline.org/`, accessed 10th May 2010

2.6 Text-mining for molecular biology researches

Text-mining or data-mining is the process of computationally deriving required information from free-form text [Cohen and Hersh, 2005][Jensen *et al.*, 2006]. The technique allows the application of algorithms, statistics and data management methods to the vast amount of literature for the identification of needed information and relationships among entities of interest [Cohen and Hersh, 2005]. New knowledge can be revealed from connecting missing relationships between information [Jensen *et al.*, 2006].

As more biomedical and ecological researches are published, the underlying knowledge of the habitats of microorganisms is expanding at an increasing rate. Text mining is a way to cope with this information overload. The technique can be used to reveal the needed knowledge that would otherwise be obscured by the large amount of information [Cohen and Hersh, 2005].

For example, the literature-mining software, Peregrine [van Haagen *et al.*, 2009], was successfully used to refer an undocumented interaction between two proteins. Enju[4] is a full text parsing tool that was used to parse 70 million sentences in MEDLINE and extracted all the biomedical entities and relationships such as protein-protein interactions. The results from this deep-parsing process were then used as a data source for a semantically aware search tool, MEDIE[5]. Similarly, GeneWays [Rzhetsky *et al.*, 2004] employs a deep parsing tool, GENIES [Friedman *et al.*, 2001], to extract biomedical knowledge of different types of binary relationships between genes, gene products, disease and drugs in signal transduction pathways [Sainani, 2008].

In the study covered in this thesis, text-mining techniques were employed to gain more information of microorganisms and their habitats or isolation sources. More details are given in Chapter 5.

2.7 Classification of microorganisms habitats

In order to compare the genetic information of different microorganisms in relation to the environments they reside in, external entities with similar physiochemical properties must be grouped together that are considerably distinct from different ecological niches. For example, mucosa-associated and non-mucosa-associated microorganisms would need to be distinguished so that sets of protein families required for microbes to interact with mucosal surfaces can be inferred.

Ontologies allow standardisation of controlled terms and the integration of different sets of terms

[4] `http://www-tsujii.is.s.u-tokyo.ac.jp/enju/index.html`, accessed 5th December 2010

[5] `http://www-tsujii.is.s.u-tokyo.ac.jp/medie/`, accessed 5th December 2010

or vocabularies. These defined terms are designed to be searchable, sharable and accessible by both humans and software agents [Lord *et al.*, 2003]. Terms may be defined related to each other in a variety of ways, for example 'is-a' and 'part-of' relations. Relations between terms may be subjected to rule-based constraints, such as the types of terms that may be related by inference. For example, if we defined that 'lake' 'is-a' 'aquatic_environment' and we know that organism_A lives in a lake. By reasoning the defined class-relation, it could be inferred that organism_A lives in an aquatic environment.

One well-known ontology used widely in the genomics research and bioinformatics communities is the GO [Ashburner *et al.*, 2000]. The GO consortium has an aim to standardising the representation of genes or gene products across species [Ashburner *et al.*, 2000]. The GO provides controlled vocabularies for biological processes, molecular functions, and cellular locations. These GO terms might be related to each others with relationship 'is-a', 'part-of' and 'regulates'[6], allowing GO terms to be structured into Directed Acyclic Graphs (DAGs).

The Open Biomedical Ontologies (OBO) foundry [Smith *et al.*, 2007] provides sets of biomedical and bioinformatics -related ontologies. One such ontology is the Environment Ontology (EnvO) [Smith and Varzi, 2002], which aims to define external entities or surrounding environments of a biological sample, such as a habitat for a microorganism. The set of habitat vocabularies is divided into several sections including host-associated or non-host associated physical material such as host body fluid, soil, marine and extreme habitats. Several EnvO terms have been reused by the Habitat-lite ontology. Habitat-lite is the first ontology that aims to create lists of terms describing habitats of organisms. All terms were selected from EnvO to form appropriate high-level terms.

2.8 Bioinformatics applications

2.8.1 Protein subcellular localisation prediction tools

Several bioinformatics tools have been designed to identify subcellular locations of amino acid sequences. A number of approaches have been employed to suit a wide variety of secretory signals commonly found in various extracytoplasmic proteins. These tools work with different cell surfaces and protein translocation mechanisms of the different taxonomic groups (described in Table 2.3 and 2.4). The prediction tools considered for this study are described here. This section is divided into two subsections: targeting signal predictors and protein subcellular localisation predictors.

[6] `http://www.geneontology.org/GO.ontology.relations.shtml`, accessed 10th May 2010

Targeting signal predictors

The presence of some targeting signals and anchoring regions on peptide sequences allow the precursor proteins to be targeted to a specific protein transport system, generally defined as targeting-signals. Some distinctive sequence patterns such as hydrophobic and high polar regions have been found to enable proteins to be anchored to the cell membrane [Alberts *et al.*, 2007]. Moreover, depending on the cell surface structure, surface proteins can be anchored to the cell surface in different ways. These conserved functional domains or motifs were described earlier in this chapter (Section 2.3.3). The trends of these amino acid residue compositions enable the prediction of protein subcellular localisation by using computational methods to identify those recognisable features. Several bioinformatics tools and algorithms have been developed to identify extracellular protein sequences based on their primary amino acid sequence data.

TMHMM

TMHMM [Krogh *et al.*, 2001] is one of the most widely-used tools to detect and locate TMH and the orientation of transmembrane proteins in the lipid-bilayer membrane. TMH can be predicted due to their distinctive patterns of hydrophobic and polar regions within the sequence, which allows pattern searching and matching. TMHMM implements its predictions through a Hidden Markov Model (HMM) algorithm.

LipoP

LipoP 1.0 [Juncker *et al.*, 2003] is another signal peptide identification tool that predicts the presence of a signal peptidase II (SPII) cleavage site found in lipoproteins, which SignalP is not capable of detecting. The tool successfully employed a HMM to distinguish SPII cleavage sites from signal peptidase I (SPI) cleavage sites. LipoP is trained with Gram-negative bacterial protein sequences from organisms belonging to the two phyla Proteobacteria and Spirochetes. However, the tool is also capable of predicting Gram-positive bacterial lipoproteins [Rahman *et al.*, 2008]. LipoP can predict lipoproteins at an accuracy rate of 96.8% and 92.9% from a test set of Gram-negative and Gram-positive bacteria, respectively.

SignalP

SignalP [Dyrlovbendtsen, 2004] is a tool for detecting a protein targeting feature on an amino acid sequence. Precursor proteins targeted to the Sec secretory pathways normally have an N-terminal signal peptide that can be detected by SignalP. These proteins are typically characterised by an N-terminal signal that can be recognised and cleaved by SPI. The tool provides two different options for running the analysis: Neural Network (NN)- and HMM- based algorithms. The tool has been trained with several bacterial and eukaryotic proteins, but not with archaeal proteins. It is reported that SignalP-HMM is more sensitive in detecting signal peptides than SignalP-NN. Conversely, SignalP-NN has a higher accuracy in predicting correct cleavage sites [Kall, 2004]. However, not every protein predicted to have a signal peptide is identified as an extracellular protein. For example, some proteins might have other retention signals, such as the ER retention signal, which holds the protein in the ER in eukaryotes.

Phobius

Phobius [Kall, 2004] utilises a HMM to combine the prediction of N-terminal signal peptide (SP) and TMH regions. This tool was developed to improve the accuracy of the discrimination of those two hydrophobic features. Determining the presence of a SP also provides the correct prediction of the transmembrane topology as it dictates that the N terminus of the mature transmembrane protein is extracytoplasmic. Since SP and TMH are highly similar, SPs were often mis-predicted as TMH by tools particularly trained to predict TMHs. When applied to the well-annotated human and *E. coli* proteomes, Phobius has proved to drastically reduce the misclassification of the two hydrophobic classes compared to SignalP and TMHMM. Phobius yielded fewer misclassifications of TMHs as SPs and *vice versa* in relation to the compared methods. However, Phobius is less sensitive than SignalP when predicting SP and cleavage sites.

Protein subcellular localisation predictors

BaCello

BaCelLo [Pierleoni *et al.*, 2006], based on decision tree of binary Support Vector Machine (SVM), has been shown to be one of the most efficient predictors for the cellular location of proteins in eukaryotes, particularly, in animals, fungi and plants. The tool implements a specific predictor for individual eukaryotic kingdoms. Proteins are divided into five classes: secreted, cytoplasmic, nuclear, mithocondrial and chloroplastic proteins. The advantage of BaCelLo over other predictors is that it considers information from the whole sequence as well as from both the N- and C-termini. It also takes extracellular-exposed sequence features into consideration. It outperforms other methods in predicting eukaryotic secreted proteins in terms of accuracy [Casadio *et al.*, 2008].

PSORTb

PSORTb [Gardy, 2004] is another well-known computational tool that uses a sequence-based SVM method for predicting bacterial protein localisation. The tool implements a BLAST homology search of proteins of known localisation. It also implements a set of analytical modules that run independently to scan for particular signals, anchoring and extracellular-exposed sequence features (all features described in Table 2). PSORTb has been reported as the most precise tool for the identification of cellular locations of both Gram-positive and -negative bacterial proteins [Gardy and Brinkman, 2006] [Gardy, 2004].

At the time of this study, the tool requires ‘root’ access in order to install successfully on Linux machine. Thus, it was not feasible to be run within a high-throughout framework, where a tool would have needed to be installed automatically prior processing.

PSORTdb

PSORTdb [7] is a web-accessible data resource for bacterial protein subcellular localisation [Rey *et al.*, 2005]. The database contains two sublocalisation databases, ePSORTdb and cPSORTdb. The former database is composed of subcellular localisations of proteins based on an experimentally verified data set. The latter database contains a pre-computed data set of proteins with their predicted subcellular localisations.

2.8.2 Detection of Protein signatures

One method to infer protein functions from a primary protein sequence is to search for known characterised features such as motifs, patterns or functional domain regions. To date, several protein signature recognition approaches have been developed to fulfil different aspects of sequence analysis resulting in many independent algorithms and databases. These resources have different strengths

[7] `http://db.psort.org`, accessed 10th May 2010

depending on the underlying analysis methods and objectives they were designed to meet. Therefore, one of the best strategies for protein sequence analysis is to combine the search results from all of these different approaches and databases.

InterProScan

InterProScan[8] [Zdobnov and Apweiler, 2001] is an application that integrates several protein signature recognition resources into one tool in order to identify previously known protein signatures. The protein signatures can be detected by InterProScan include protein domains, families, and functional sites deposited in the InterPro member databases. These member databases include SUPERFAMILY [Wilson *et al.*, 2009], PROSITE [Hulo *et al.*, 2006], Pfam [Finn *et al.*, 2010], PRINTS [Attwood, 2002], ProDom [Corpet *et al.*, 1999], PIR-PSD [Barker *et al.*, 1999], SMART [Schultz *et al.*, 2000], TIGRFAMs [Haft *et al.*, 2001] and HAMAP [Lima *et al.*, 2009]. InterProScan also provides a function to look up corresponding InterPro [Hunter *et al.*, 2009] and Gene Ontology [Ashburner *et al.*, 2000] annotations on a given protein sequence. The tool is available in both a web-based form and a download for a local installation. InterProScan is a Perl-based program that chains together other existing protein signature recognisers and relevant tools such as HMMER, PatternScan, ProfileScan, FPrintScan and gapped-BLASTP.

HMMER

HMMER[9] [Eddy, 1998] is a software package for protein sequence analysis. The software includes several functions built on the basis of HMM protein profiles. HMMER provides functions for: constructing an HMM-profile from a sequence alignment; searching a sequence or a sequence database for particular HMM-profiles. The software is used as a core utility in the well-known public protein family databases such as Pfam [Finn *et al.*, 2010] and InterPro [Hunter *et al.*, 2009].

2.8.3 Profile-profile comparisons

Protein sequences sharing $< 30\%$ identity tend to have significant functional differences [Todd *et al.*, 2001]. However, the structure of the proteins can still be inferred for very distantly related proteins [Zheng *et al.*, 2005]. Therefore, for remotely homologous proteins, the structures and folds of a pro-

[8] http://www.ebi.ac.uk/Tools/InterProScan/, accessed 10th May 2010

[9] http://hmmer.wustl.edu/, accessed 10th May 2010

tein can be used to infer of their structural models, potential active sites and substrate binding regions [Söding *et al.*, 2005].

HHpred server for remote homolog detection and 3D structure prediction

HHpred[10] is a tool for protein sequence homology detection and structure prediction based on the pairwise comparison of protein HMM profiles [Soding, 2005]. While most conventional sequence search methods search sequence databases such as UniProt or the NR, HHpred searches alignment databases such as Pfam, InterPro or SMART. The use of the HMM-HMM comparisons provides sensitive results for finding remote homologs. The tool is easy to use and is faster than other protein structure prediction servers (e.g. Profile Comparer, COMPASS and PROF_SIM).

2.8.4 High-throughput data analysis in bioinformatics

The last decade has seen a rapid increase in the number of completely sequenced genomes. This tremendous amount of sequence data is flooding into genome databases (see Section 2.1), necessitating the development of efficient tools for comparative genome sequence analysis. To utilise the wealth of genomics data, a high-throughput computational framework is required to support sequence analyses and workflow enactment [Ahmed, 2009]. Grid and Cloud computing technologies permit large numbers of computers to be used in parallel [Foster *et al.*, 2001] [Andrade *et al.*, 2006] [Flanagan, 2009].

Grid and cloud computing

Grid computing permits multiple institutions to combine and share their computing resources [Foster *et al.*, 2001][Baker *et al.*, 2002]. Users at a different institution are able to migrate their computational work to take advantage of the spare capacity available at another institution [Frey *et al.*, 2002]. However, Grids are typically difficult to set up and maintain [Ibrahim *et al.*, 2008]. In particular, security concerns often limit the ability of remote users to install domain-specific software [McNab, 2003]. Users are typically given a restricted account that prohibits them from machine administration and limits their resource usage [McNab, 2003].

In contrast to Grid computing, Cloud computing utilises remote computational power and is provided on a commercial basis. Cloud computing providers generally do not provide access to the physical

[10] `http://toolkit.tuebingen.mpg.de/hhpred`, accessed 10th May 2010

hardware, rather, end-user software is executed within one or more virtual machines [Xu, 2010] [Smith and Nair, 2005]. Virtual machines (VMs) are software processes that present a virtualised hardware environment to applications [Smith and Nair, 2005].

Several VMs may execute at simultaneously on the same physical machine. The use of VMs offers a number of advantages to both the provider and the user. For example, providers may offer numerous configurations of virtual machines with different specifications in terms of number of CPUs or RAM. Security risks are also migrated through the use of isolated VMs. For example, access to other users' processes or files are not possible. Moreover, VMs allow users to have complete control over their environment, permitting specialised software to be installed with no restrictions. Another advantage of Cloud computing is the ability to expand and reduce the number of rented CPUs or storage capacity as required [Xu, 2010].

Both Grid and Cloud technologies allow execution of high-throughput parallel computational tasks over distributed computers on a network. As the size and complexity of a distributed computing system increases, there is an increasing requirement for automated management systems to assist when inevitable hardware or software component failures occur. Many Grid middleware implementations such as Microbase [Wipat *et al.*, 2004][Sun *et al.*, 2005] and Globus [Foster and Kesselman, 1997] [Foster, 2006] provide useful job management features such as notification services, workflow enactment, resource discovery and provenance tracking.

Bioinformatics workflow

Complete genome sequence data are released rapidly and ever more genome sequencing projects are getting underway. Comparative genomics of large-scale data sets across taxonomic groups facilitates a better understanding of the structural diversity and evolutionary origin of proteome from various perspectives. To establish a set of putative extracellular proteins from different taxonomic groups, a number of approaches or tools would be required to predict a wide variety of secretory targeting signals, transmembrane regions and other well-characterised extracytoplasmic protein signatures. Such a problem can be addressed using an e-Science approach where a computational infrastructure is used to aid the integration of heterogeneous data sets or software through scientific workflows implemented across a distributed computing framework [Craddock *et al.*, 2008][Ahmed, 2009]. Workflow is an approach that allows connections between a set of different execution units. Various tools can be chained together and executed orderly one after another. Output from one step can be parsed and passed to another step automatically.

However, several issues must be considered when multiple independent tools are combined for an efficient workflow. These issues include data compatibility between programs, and the computational and logistical requirements of executing standalone programs or multiple instances of programs at each workflow step.

Several workflow construction tools have been developed considering the issues above, such as, Taverna. Taverna provides an alternative to 'cut and paste' content integration between bioinformatics analysis websites, and reduces fragile screen-scraping integration scripts [Hull *et al.*, 2006].

Taverna

Taverna[11] [Hull *et al.*, 2006] is an application that provides a one point service for constructing and running bioinformatics workflows. The application makes use of Web Service
[Neerincx and Leunissen, 2005] to enable the integration of programs and data sources. Web services are an accepted industry standard that permits well-defined programmatic access to data sources [Neerincx and Leunissen, 2005]. The services provided by autonomous third parties can be programatically accessed over the network via Web Services. Therefore, the selected tools and databases do not need to be installed locally on the user machine. Taverna also provides a Graphical User Interface (GUI) to facilitate constructing and enacting workflows, as well as browsing the output of workflows.

Microbase

Microbase [Flanagan, 2009][Wipat *et al.*, 2004] is an event-driven, service-oriented, Grid system capable of executing analysis pipelines automatically. Microbase provides a modular framework that facilitates the development of applications, allowing them to utilise Grid and Cloud resources. This component-based computational system is designed to provide an environment for analysing large-scale data in a high-throughput, distributed fashion. The system enables complex analysis pipelines consisting of multiple analysis tools to be constructed. The framework also allows efficient automation of various steps of the analysis processes involving a research study, facilitating systematic analyses.

Small-scale bioinformatics analyses may be performed manually in an ad-hoc manner by a researcher using a single desktop computer. However, as the amount of data needing to be analysed increases,

[11] http://www.taverna.org.uk/, accessed 20th July 2010

an automated, systematic approach becomes more desirable. The Microbase system provides programmers with a means to construct highly parallel analyses that can execute within Grid and Cloud environments.

Microbase applications consist of one or more modules, termed ‘responders’. A responder consists of two parts: a server-based program and a mobile program capable of migrating between computers. The server-based part is typically a Web Service and is responsible for task scheduling and data management operations. For example, the server-based component may maintain a relational database and service queries for data from other responders. The mobile program is responsible for implementing computationally-intensive operations associated with generating the data set managed by a particular responder, such as executing a bioinformatics application. Multiple copies of these ‘compute jobs’ may exist on multiple machines simultaneously, permitting a large computational task to be performed in parallel [Flanagan, 2009].

Microbase consists of four main components [Flanagan, 2009]:

- A set of responders to perform the computational work and data management functions required by an analysis step.
- An event-based notification system to co-ordinate a set of responders.
- A job server that matches jobs to available machines. The job server also handles job failures if they occur, and re-schedules failed jobs.
- A resource system stores output files produced by bioinformatics applications. The resource system is also responsible for distributing program and data files to computers as necessary.

The Microbase architecture is shown in Figure 2.9. Each component will now be discussed in more detail.

Microbase notification system

The notification service is responsible for informing responders of new events as they occur. For example, an event could be created as a result of a new genome sequence becoming available for analysis, or that a particular bioinformatics analysis task has completed. Therefore, a notification event can be used as the trigger to start a set of computational work. The completion of that work may then may be reported to other responders as a new notification event, which might trigger further analyses to run. This approach allows for processing pipelines to be constructed that are composed of a number of responders (see Section 2.8.4).

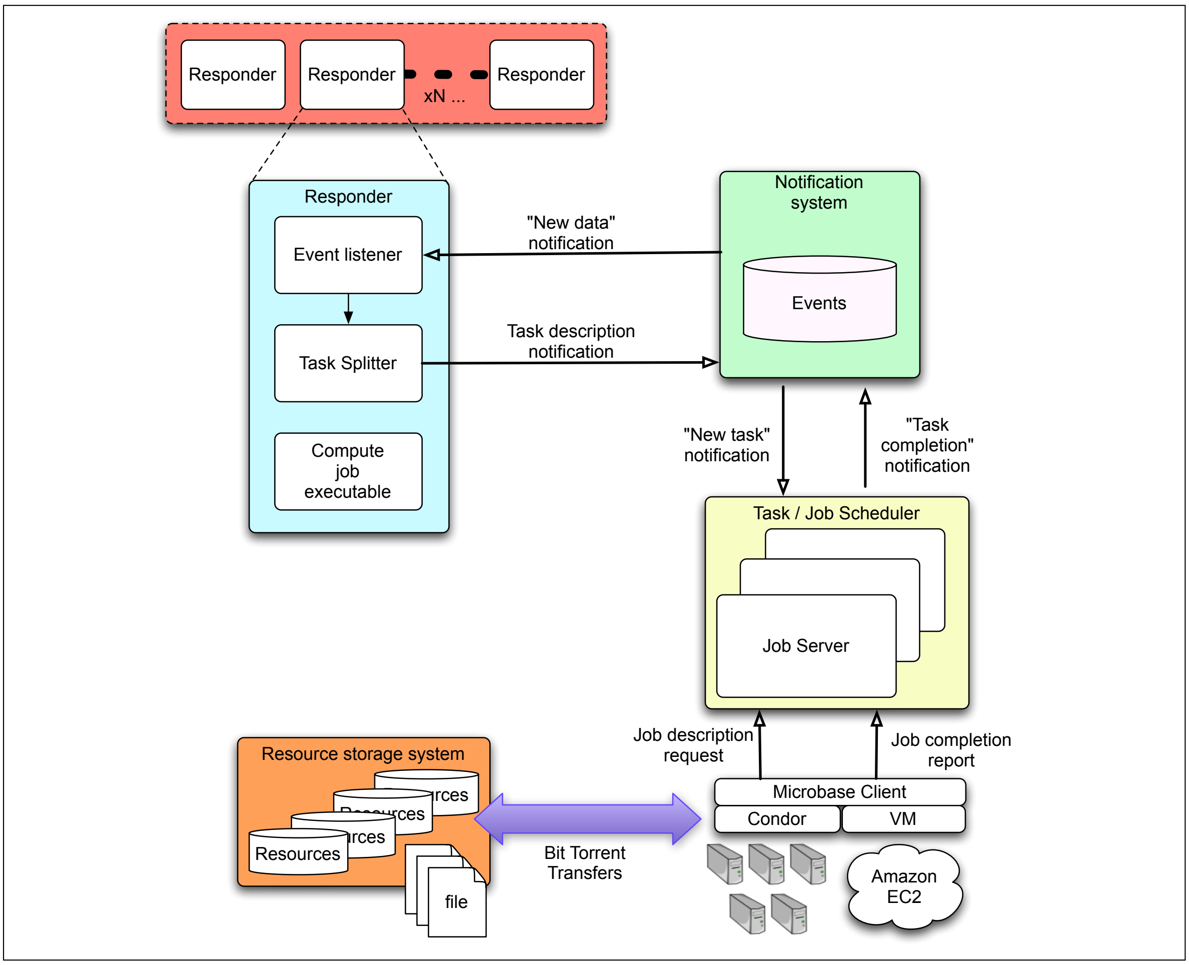


Figure 2.9: Overview of the Microbase architecture. The notification system co-ordinates all system processes, including the core Microbase services as well as user components. Users may add their own domain-specific functionality to the system via components termed responders. Responders react to notification events they are subscribed to. For example, if a BLAST or InterProScan responder receives a 'new genome available' message, then they will react by requesting that the appropriate computation is performed. This is achieved by sending a 'task description' message, which is the forwarded to a Microbase job server. The server will then assign an appropriate number of Grid or Cloud machines to complete the work. Program executable files and data files are transferred to Grid and Cloud machines from the Microbase resource storage system. The file transfer uses the BitTorrent protocol to efficiently transfer large files such as multi-gigabyte blast databases. For example, a large cluster of machines can have a dynamically installed blast database in just a few minutes. (Figure adapted from [Flanagan, 2009])

Microbase responders

Microbase responders perform computational work and data management functions required by bioinformatics tools. A responder is an application-specific management component. Each responder wraps all the functionalities needed for a particular process including: an event listener for new notifications; a compute job; and task splitter and distributor. The key concept of having such a modular user-specific component like a responder is to allow a dynamic workflow to be formed and to enable modification of the structure of that workflow over time.

A responder is executed corresponding to relevant notification events. Responders can communicate with each other via an event-based notification system. Messages sent by a responder are typically used to inform other responders of new data available to the system from an external source, or from the completion of a processing operation. These messages can therefore be used to co-ordinate multiple responders, permitting complex workflows to be formed.

A responder is typically responsible for handling the needs of a single analysis application. In order to execute multiple analysis tools, multiple responders would need to be written. The set of independent responders can be co-ordinated via event notifications to form the analysis pipeline.

Each responder contains a Web Service, allowing generated data to be exposed to other responders, or to remote users and machines, located anywhere on the Internet if necessary.

Microbase job server

Job server component schedules and tracks jobs requested by a responder. Each job is implemented by available machines in the computing environment, which are efficiently identified through the Microbase job enactor. The job server provides feedback such as job statuses. The job server may relaunch jobs if they fail, or sends out a notifications when jobs successfully complete.

Microbase resource system

The Microbase resource system is a permanent archive for software and data items, as well as a scalable content distribution system. Every item is stored with a unique identifier, as well as a set of tags that facilitate searching for content. For instance, a typical query may need to locate a particular software package, with a certain version number for a particular operating system. The resource system is designed to store every output file produced at each stage of a workflow. The developer of Microbase believed that it was necessary to store each version of each application or data item for consistency and analysis repeatability reasons. The resource system is scalable, utilising a peer-to-peer transfer protocol based on BitTorrent [Cohen, 2003] transfers.

2.9 Statistical analyses to correlate phenotype to genotype

2.9.1 Univariate analysis

The pairwise significance test is one of techniques used to find whether there is a statistically significant difference between two groups and that this difference is not likely to occur by random chance alone. The Chi-square test is typically used to compare two independent categorical variables. For example, to test whether a gene or protein domain is overrepresented in organisms from a given environment, the two independent variables here would be a summary profile of the gene from different sets of organisms from a given environment versus those from other environments.

Hypergeometric distribution

The hypergeometric test is a discrete univariate probability distribution. It is a statistical significance test. The technique is similar to the chi-squared test for hypothesis test, but is more accurate for small numbers (<6) [Lozupone *et al.*, 2006]. The equation for the hypergeometric distribution is:

$$p(i \geq m \mid N, M, n) = \sum_{i=m}^{n} \frac{\binom{M}{i}\binom{M-n}{n-i}}{\binom{N}{n}}$$

As an example, to determine the probability of finding a genotypic feature in an organism with an phenotypic feature by chance: N is defined as the total number of organisms and n is the number of organisms with a given genotypic feature. M is the number of organisms expressing the phenotype, and m is the number of organism that have both the genotypic feature and also express the phenotype. The function provides the probability of finding a protein signature or domain in an organism by chance.

To identify the direction of the association, the mean value (μ) of the hypergeometric distribution can be used as a reference. The mean value can be calculated by:

$$\mu = n * M / N$$

where n, M, N and m can be referred from the previous equation. The relationship of the two variables is a positive correlation when m is bigger than μ. On the other hand, the relationship is negative or corresponds to an anti-correlation if m is smaller than μ [Liu *et al.*, 2006].

Propensity score

Propensity score Φ [Little and Rubin, 2000] is the probability of a unit in being assigned to a particular condition given a set of known covariates.

The propensity score which is referred to in this thesis was obtained from [Jim, 2003]:

$$\Phi_f(i) = \frac{\text{fraction of genomes with phenotype } f \text{ that contain protein } i}{\text{fraction of genomes that contain protein } i} = \frac{t_{i,f}/T_f}{n_i/N}$$

where T_f is the number of genomes that exhibit a phenotype f, N is the total number of genomes, $t_{i,f}$ is the number of genomes that both exhibit phenotype f and contain protein domain i, and n_i is the total number of genomes that contain protein domain i.

2.9.2 Bivariate analysis

Bivariate analysis is a statistical technique that measures two variables at a time. Correlation is an example of bivariate analysis which finds the strength of an association between two variables.

Pearson's correlation coefficient

Pearson's correlation coefficient is a widely used statistical technique to measure the degree of a linear relationship between related variables. It is one of the most common statistical measures of correlation and most successfully used method for finding genotype-phenotype association.

The formula of Pearson's correlation coefficient (r) is defined as:

$$r_{ik} = \frac{\sum_{j=1}^{N} (X_{ij} - \bar{X}_i)(Y_{jk} - \bar{Y}_k)}{\sqrt{\sum_{j=1}^{N} (X_{ij} - \bar{X}_i)^2} \sqrt{\sum_{j=1}^{N} (Y_{jk} - \bar{Y}_k)^2}}$$

where r_{ik} is the correlation strength between i and k. $\bar{Y}$ and $\bar{X}$ are the sample means of X and Y. r_{ik} ranges from +1 to -1. The closer the correlation is to either +1 or -1, the stronger the positive and negative relationships, respectively. A value of r_{ik} near to 0 means that no correlation exists between the two variables i and k.

Mutual information

Mutual information is a another statistical technique used to estimate the association between two random variables by measuring the mutual dependence of the two variables. The unit of this measurement technique is naturally normalised between 0 and 1 bits. The lower the bit value, the less information dependency between the two variables, whereas a high bit value implies a strong association between the variables. The Mutual information can be applied to both continuous values or discrete values [Slonim *et al.*, 2006]. Slonim *et al.* (2006) use the empirical mutual information to estimate mutual information between genes and phenotypes. As an example of estimating mutual information between a protein domain and a phenotype of interest: N is an 2 x 2 count matrix defined by given that a gene phylogenetic profile and a phenotype profile are known. $N(1,1)$ is the number of taxa with phenotype of interest and the protein domain. $N(1,2)$ is the number of taxa with the phenotype but without the protein domain. $N(2,1)$ is the number of taxa without the phenotype but with the protein domain. $N(2,2)$ is the number of taxa without both the phenotype and the protein domain. The empirical mutual information between the two variables is [Slonim *et al.*, 2006][Cover and Thomas, 1991]:

$$I(X;Y) = \sum_{y \in Y, x \in X} p(x,y) \log(\frac{p(x,y)}{p(x)p(y)})$$

where $p(x,y) = N(x,y)/sum_{x,y}N(x,y)$, $p(x) = p(x,1) + p(x,2)$, and $p(y) = p(1,y) + p(2,y)$.

Chapter 3

Development of a High-Throughput Sequence Analysis Workflow

3.1 Introduction

Genome databases such as GenBank and UniProt have been growing at an exponential rate for the last few years. The existing large number of complete genome sequences frequently requires researchers to automate sequence data analyses in a systematic manner. The project described in this thesis constitutes a large-scale comparative genomics study including approximately 3 million protein sequences from more than 800 organisms whose complete genomes were available at the beginning of the project. Genome information from all three domains of cellular life including bacteria, archaea and microbial eukaryotes are of interest in this study. Analysis at this scale requires an automated, systematic approach in order to be feasible [Riley *et al.*, 2007] [Decker *et al.*, 2001] [Walter *et al.*, 2009].

There are typically three considerations in the design of a high-throughput analysis: the ability to co-ordinate multiple analysis tools and data flow between tools; high level data management to ensure the completeness of result databases; and the overall computational speed. The establishment of analysis workflows enables a sequence of independent software packages to be chained together [Hull *et al.*, 2006]. Many approaches have been developed to reduce the computational time of large-scale data of analysis tasks [Foster *et al.*, 2001][Frey *et al.*, 2002][Xu, 2010]. In order to execute all of the required computational tasks within an acceptable time frame, highly parallel computational techniques such as Grids [Foster *et al.*, 2001] or Clouds [Xu, 2010] are often employed [Karo *et al.*, 2001] [Walter *et al.*, 2009] [Matsunaga *et al.*, 2009]. These technologies allow computationally-intensive work to be distributed and shared amongst a cluster of computers.

Several existing bioinformatics tools were required to facilitate the investigation of the various features of protein sequences. As a result, the project necessitated large amounts of computational time to analyse the very large input data sets involved. Several workflows were developed to perform these analyses utilising Grid and Cloud computing technologies, and to integrate the resulting data. Workflows used to orchestrate several bioinformatics tools on distributed-computing systems are described in this chapter. In later chapters (Chapter 4 and Chapter 6), the resulting integrated data sets are analysed in order to formulate biological hypotheses.

3.1.1 Objectives

One of the main aims of the work presented in this thesis was to analyse a large number of protein sequences for the identification of features facilitating the mucosal lifestyle of microorganisms (see Section 1.2, 1.3). The computational phase of the project involved the implementation of a bioinformatics analysis workflow using a distributed-computing system. The objectives covered by this chapter facilitate this aim by addressing the practical and logistical challenges associated with building and maintaining large heterogeneous data sets. The objectives covered by this chapter are:

- to design and implement a bioinformatics workflow that combines various computational analysis processes in order to automate a large-scale processing of sequence data for the identification of extracytoplasmic proteome and sequence features;
- to execute the workflow by using Cloud and Grid computing technologies in order to investigate their suitability for large-scale bioinformatics analyses;
- to construct a set of interconnected databases in order to allow new knowledge to be extracted from the raw workflow output data. These databases will be used as a foundation from which further statistical analyses will be performed (discussed further in chapters 4 and 6).

3.2 Methods

A distributed computational framework called Microbase (see section 2.8.4) was employed to construct a bioinformatics framework for genome data analysis. Microbase fulfilled the requirements of this project since it facilitates the construction of bioinformatics workflows as well as providing a distributed computing environment for processing multiple computational tasks in parallel. The framework allows the use of idle desktop computers as well as dedicated servers for data processing. Moreover, Microbase allows the use of relational databases as a structured data store to store outputs produced by an encapsulated bioinformatics tool. Therefore, Microbase was a highly suitable framework for the large-scale analyses required by this project.

More than 3 million protein sequences were analysed by six different bioinformatics tools including: TMHMM, SignalP, LipoP, InterproScan, BLASTP and HMMER. A relatively large amount of result data was expected to be generated from the high-throughput computational workflows. It was therefore necessary to establish an efficient and organised data storing process. Relational databases were used to store the results from each tool; one database per tool.

To construct a bioinformatics workflow, the project utilised a key functionality provided by Microbase, the event-based notification system (see Section 2.8.4). The workflow developer creates components, called responders. Each responder encapsulates a user-specific functionality, for example, bioinformatics tool such as BLAST, HMMER) (see Section 2.8.4). Each responder is registered with the notification system such that it activates upon notification of specific event(s). A responder is only triggered by the specific types of notification message(s) that it is registered to receive. A message published by one responder can activate one or more responders that have registered their interest in that event. As a result, an automated workflow can be established permitting data to flow from one responder to the next. In the rest of this chapter, the following issues are discussed:

- an overview of the project's sequence analysis workflow;
- a detailed discussion of the responders developed to perform various analysis processes and the bioinformatics tools they encapsulated;
- an introduction to the relational database for storing the analysis results produced by the workflow, and its role in providing biological knowledge from the integrated result sets.

The Microbase system was deployed on Newcastle University servers as well as ordinary university cluster room PCs. Amazon EC2 Cloud[1] resources were also used to execute automated high-throughput scientific workflows in a systematic manner.

3.2.1 Overview of the sequence analysis workflow

A series of project-specific responders were developed to encapsulate several bioinformatics tools for performing various sequence analysis tasks (Figure 3.1). The overall project workflow was designed

[1] http://aws.amazon.com/ec2/, accessed 15th December 2010

to carry out a number of functions, including

- The automatic retrieval of sequence data from a public genome resource (NCBI [2]) and generation of a set of genomics input data for the downstream analysis processes.
- Processing the input sequence through various bioinformatics tools in order to predict protein subcellular localisations, and to recognise protein domains. Protein similarity searches were also performed.
- Extraction and transformation of relevant output from bioinformatics tools to feed into the later stages of manual analysis involving various statistical approaches.

3.2.2 Design pattern for project-specific responders

Every responder developed for this project is comprised of two main components: a server-based data management component that includes a Web Service and database; and a distributable computational unit that executes on multiple worker machines.

The server-based data management component runs on dedicated hardware. This component responds to notification events which initiate the scheduling of computationally-intensive jobs [Flanagan, 2009]. A 'job' represents an independent unit of computational work, such as a single BLAST command line execution, to be distributed to and processed by a single worker machine. A single notification event may result in the scheduling of many jobs. For example, the addition of a single new genome sequence might trigger hundreds of BLAST executions in order to add to an all-against-all comparison data set. In addition to job scheduling, the server-based component is also responsible for the management of a responder-specific structured data store used to store the output of its associated bioinformatics tool. Microbase employs Web services to mediate access to the structured database, enabling analysis results to be parsed directly into the data storage for future use [Flanagan, 2009].

3.2.3 Primary data acquisition and storage

Much bioinformatics data is made available in the form of free-form, or semi-structured text files. These file are straightforward for human to read, but are inefficient for computers to query. In order to extract the information stored in a flat file for storage in a structured database, the flat file must

[2] www.ncbi.nlm.nih.gov, accessed 10th August 2010

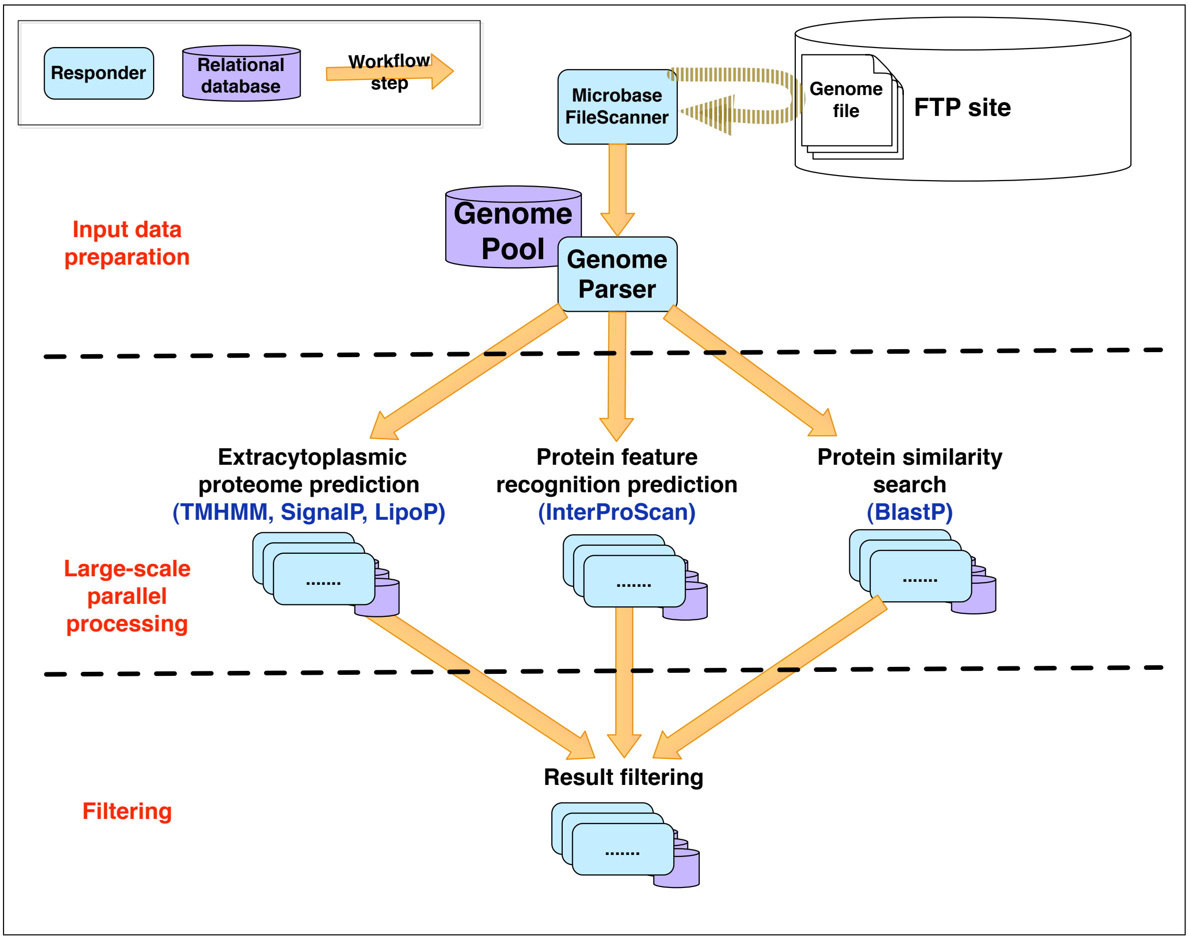


Figure 3.1: Summary of the protein sequence analysis workflow implemented as a set of Microbase responders. Project-specific Microbase responders are presented as blue squares. All responders communicate with their associated relational databases (purple cylinders) via their respective Web Service. The overall workflow is divided into three stages: (1) input data preparation; (2) large-scale parallel processing; (3) result filtering. The first stage of the workflow involves the Microbase 'FileScanner' and the 'GenomeParser' responders. These responders perform an automated retrieval of data from an FTP site, extract and parse the input data into a relational database called the GenomePool (discussed in section 3.2.4). The second stage can be divided conceptually into three pipelines for sequence data analysis. The responders in this stage encapsulate various bioinformatics tools, such as TMHMM, InterProScan, and BLASTP, to perform different logically-related and complementary analyses of large-scale input sequence data. The final stage is then invoked to extract the relevant results for further analysis steps. These results are produced from the three pipelines in the previous stage. An example of the result-filtering database, called PredExtDB, is used to store a list of predicted extracytoplasmic proteins and their corresponding information (see Section 3.2.4). Each responder developed for the workflow is described in more detail in Sections 3.2.3, 3.3 and 3.3.3.

be processed by an appropriate parser. GenBank [Benson *et al.*, 2009] provides a large repository of genome information. Much of this data is made available through files on NCBI's FTP site. However, the information contained within GenBank-formatted files is difficult to obtain due to the plain-text formatting (flat-file structure). For example, if the translation of a particular coding region was required, each file would need to be scanned until a match was found. Therefore, there is a need to extract data from the text file and transform the information into a structured form, enabling efficient data querying. In this section, the processes of obtaining genome data files and reformatting the data using the Microbase-provided and project-specific responders are described. These processes are termed 'FileScanner' and 'GenomeParser', respectively.

Genome data acquisition

In order to obtain genome data files from an FTP site, an existing Microbase responder called the FileScanner responder was utilised. FTP[3] is a protocol allowing computers to transfer files over a network, such as the Internet. The FileScanner responder is responsible for detecting the arrival of a new file on an FTP site and for passing the file into the Microbase resource system. In this project, the FileScanner responder was configured to search for GenBank-formatted (gbk) files on a local FTP site. The local FTP site holds GenBank files taken from the Reference Sequence (RefSeq) [Pruitt *et al.*, 2009] NCBI FTP site [4,5,6] (accessed 11 February 2009). Genome fragments acquired for use by this project include complete genome sequences of bacteria and archaeal chromosomes, plasmids, and eukaryotic chromosomes and their organelle genomes (if available). RefSeq provides a non-redundant, curated set of sequences for transcripts, proteins and genomics DNAs [Pruitt *et al.*, 2005][Pruitt *et al.*, 2009].

The complete or draft genome sequence data for some known mucosa-thriving eukaryotic microorganisms were not available on the NCBI FTP site. The GenBank-formatted files for these organisms were missing. Organisms with missing data files included: *Entamoeba histolytica* HM-1:IMSS, *E. dispar* SAW760, *Giardia lamblia* ATCC 50803, *Trichomonas vaginalis G3*, *Cryptococcus neoformans* var. neoformans B-3501A, *Coccidioides immitis* RS, *Aspergillus terreus* NIH2624, *A. clavatus* NRRL 1, *Leishmania major* strain Friedlin and *L. braziliensis* MHOM/BR/75/M2904. These genome sequences were only accessible via the NCBI Web interface which is not ideal for large-scale data retrieval due to its unreliability. An additional script was then developed in order to iteratively

[3] `http://www.faqs.org/rfcs/rfc959.html`, accessed 20th November 2010
[4] `ftp://ftp.ncbi.nih.gov/genomes/Bacteria`, accessed 11st February 2009
[5] `ftp://ftp.ncbi.nih.gov/genomes/Fungi`, accessed 11st February 2009
[6] `ftp://ftp.ncbi.nih.gov/genomes/Protozoa`, accessed 11st February 2009

retrieve genome information from the web interface using Entrez Programming Utilities (E-utilities)[7] to complement the bulk of the data available in GenBank files. More than 100 records of contigs or scaffolds were available for some organisms, so an additional step was required to merge those fragments into fewer number of files to reduce the workload in the analysis workflow. For example, the protist *T. vaginalis* G3's draft genome sequence encodes approximately 59,000 protein sequences and is one of the most difficult to assemble due to genome repetition by a large number of massive gene duplications [Carlton *et al.*, 2007]. More than 20,000 files (contigs) were found to be derived from the *T. vaginalis* genome. A smaller set containing 17 manually concatenated files in GenBank format were then generated. These 17 GenBank-formatted files represent the 17,290 contigs.

The GenBank files downloaded from the FTP site were pooled together with the files generated through querying the Web interface. For convenience, these files were placed into a locally-hosted FTP site so that the Microbase FileScanner responder could be used without modification. The files in this local FTP site were then detected by the FileScanner. For each GenBank file found, an event notification message was fired. The downstream responders were then notified of the availity of new data. In this project, messages from the FilScanner responder were configured to activate the GenomeParser responder (more detail see Section 3.2.3).

Data extraction by GenomeParser Responder

The GenomeParser responder is responsible for extracting genome information from the plain text files and storing it within an indexed, structured database for convenient access from other responders or users. A structured in-house database, called the GenomePool, was developed to store all of the genome information processed by this responder (see Section 3.2.4). A GenBank-formatted genome file contains genome sequence information required for the project (details listed below). For this project, protein sequence information is the centre of interest. The content in a GenBank-formatted (.gbk) file is structured in a way that is easily readable for humans. The information is also programmatically accessible with an appropriate parser.

The function of the GenomeParser responder is to await 'new file' event notifications received from the FileScanner responder. When a new GenBank file is detected, the GenomeParser schedules a compute job to run, which is responsible for parsing the plain text GenBank file and inserting the contents into the GenomePool database (see section 3.2.4). Multiple files may be parsed at the same time, allowing a degree of parallelisation. In addition to maintaining a structured data store,

[7] `http://www.ncbi.nlm.nih.gov/books/NBK25501/`, accessed 20th October 2010

the GenomeParser also generates two FASTA-formatted files for each GenBank file parsed: one containing the whole genomic DNA sequence of a given genome file, and another containing gene-coded protein sequences. The genomic DNA sequence was extracted from the 'ORIGIN' section in the GenBank-formatted file, while protein sequences were extracted from 'translation' tagged lines within 'CDS' sections. However, if an amino acid sequence on the 'translation' line was absent from the GenBank file, then a RefSeq accession number for that gene product sequence was used as a query to automatically fetch a corresponding gene-coded protein sequence from the Web interface to RefSeq database. The generated FASTA files were stored in the Microbase resource system, ready to be used by other responders.

Data set extraction

The GenomeParser parses GenBank-formatted files, and extract information relevant to this project, which is then inserted into the GenomePool database. The information extracted is listed below:

- Metadata of a genome sequence: RefSeq accession number, version number.
- Taxonomic information: a scientific name of a source organism with a corresponding NCBI taxon identifier and taxonomic lineage.
- Coding sequence information: start and stop codons, gene names, locus tags, gene ids, gene-coded protein accessions and annotations;
- Sequence data: amino acid sequences of gene products, and nucleotide sequence(s) of the genomic DNA for a given genome fragment.

A GenBank-formatted genome file normally contains one genomics element. This genome sequence could be a complete plasmid genome, a complete chromosome, an eukaryotic organelle's genome, or a contig or scaffold of a draft genome. The data representing a single organism can therefore be spread across more than one GenBank file. For example, two GenBank files exist for to *Bacteroides fragilis* NCTC 9343: one is the complete genomic DNA sequence and another is a plasmid DNA sequence.

GenomeParser notification message contents

Each time a GenBank file has been successfully parsed by the GenomeParser, a 'new genome available' notification event is published by the GenomeParser responder. This message is received simultaneously by registered downstream analysis responders (see section 3.3). The content of a 'new genome available' message consists of information from the GenBank file as well as metadata from the GenomePool database that may be used by responders to trigger further down stream analysis. The GenBank-extracted information provides biological meaning and standard references for a genome fragment, whereas the GenomePool-generated metadata is useful as an internal reference for a given sequence for use by the Microbase system or intermediate databases. For example, the message content includes a RefSeq accession number which is a standard identifier and can be used to gain more information about a sequence file from the public NCBI database. The message also contains a GenomePool-generated protein-FASTA file identifier which is used by several Microbase components as a reference to a particular FASTA-formatted file of protein sequences. The complete content of a GenomePool notification message are as follows:

Message contents from a GenBank record

The contents listed in this section were obtained directly from the GenBank-formatted genome file.

- RefSeq accession[8] and version number[9]: this accession is provided by NCBI as a unique identifier for each genome sequence in the RefSeq database. A version number indicates the current revision of the file. A new version number is assigned by NCBI once a new set of annotations were added or any change to the sequence data was made. These identifiers were extracted from a GenBank-formatted file where lines were tagged with `ACCESSION` and `VERSION`, respectively.

- Taxon identifier: a reference number that specifies the taxonomic ranking of a given organism in the NCBI taxonomy database.

- Organism name and taxonomic linage: this name represents a scientific name of an organism from which the genome was derived. The taxonomic lineage provides a summary of the evolutionary origin of the organism. This information was extracted from lines tagged with `ORGANISM` in a GenBank file.

- Genome description: this field was extracted from the `DEFINITION` tagged line from a GenBank-formatted file. The field provides a brief textual description of the genome sequence, including information such as source organisms, sequence names and a human-readable description of the sequence's functions.

[8] http://www.ncbi.nlm.nih.gov/refseq/key.html#accessions, accessed 20th October 2010

[9] http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html, accessed 20th October 2010

- Fragment type: this field was extracted from a GenBank file from the FEATURES section. The type of genome fragment is usually noted within a subsection called ‘source’. By an observation through all GenBank genomes files obtained for this project, genome fragment types can be classified into the following category: genome; plasmid; chromosome; organelle; and unknown. The ‘unknown’ type was assigned where no information of a genome fragment type was provided in the file.

Message contents from the GenomePool database

In addition to data extracted from a GenBank file, the GenomrParser notification message also contains the following fields from the GenomePool database:

- GenomePool fragment identifier: this identifier is generated by the GenomePool database once a genome file has been stored. This fragment identifier can be considered as an in-house unique identifier for a given genome file.
- Organism identifier: a reference number generated by the GenomePool database to be used as an unique identifier of an organism whose genome sequences were stored in the in-house database.
- File identifier: an identifier assigned by the Microbase FileScanner responder to each genome file that was detected by the responder. Each identifier is therefore associated with an actual file that has been deposited into the Microbase resource system.
- DNA FASTA file identifier: an identifier assigned by the GenomeParser responder to a FASTA file containing the whole genomics DNA sequence for a given GenBank genome file. The file was generated during the parsing process.
- Protein FASTA file identifier: an identifier assigned by the GenomeParser responder to a FASTA file containing the gene-coded amino acid sequences derived from the genome. These identifiers are used by downstream responders to retrieve a collection of protein sequences for various sequence analyses.

3.2.4 GenomePool and analysis result databases

Several in-house relational databases were developed to deposit biological data produced by the GenomeParser responder (see Section 3.2.3) and to store results from every sequence analysis tool used in the project. This section provides a detailed description of the main functionalities and properties of the project's primary, secondary and tertiary databases which are populated by a series of project-specific Microbase responders (see Section 3.3 for more details). The GenomePool is an in-house primary database that was developed to store input sequence data to be analysed. Secondary databases are used for storing the output produced from the project's high-throughput analysis workflow (see Section 3.3). Finally, a database of predicted extracytoplasmic proteins is a tertiary data store consisting of a selection of entries from the secondary databases. In this section, the primary and the tertiary databases are described in detail, while the specific secondary databases are described as necessary throughout the rest of the chapter.

GenomePool database

The GenomePool database (GPDB) is a structured relational database designed to be used as an in-house repository of sequences and their annotations derived from public genome sequence databases. Publicly available data files are read and information useful to this project is extracted and parsed into the GPDB via the GenomeParser responder (see Section 3.2.3). The GPDB is a back-end data storage for the GenomeParser responder. The GPDB has fields to store information contained in genome sequence files including the actual sequence data, associated metadata and annotations. Data stored in the GPDB covers most of the information in the genome file including: locus tag, accession number, version number, source organism, genome source (chromosome, plasmid, organelles), descriptions of the genome fragment, the annotation of genes and proteins, and the actual sequence data. In addition, the GPDB also stores all available information about genes and gene products including: regions of biological significance and their annotations such as start and stop coding sequences, gene orientation, gene name, locus tag, gene product, protein identifier and other sequence features annotated on the protein sequence. In this project, the GPDB stores sequence information read from GenBank-formatted genome sequence data files, obtained from the RefSeq database.

The GPDB acts as a central database that links phenotypic information of microorganisms to various aspects of their genotypic features predicted by various analysis processes. The GPDB may be used either as a standalone repository of genome information, or as part of a larger pipeline. When used within a processing pipeline, other pipeline components may query the GPDB via its Web Service interface.

Secondary databases

The secondary databases are responsible for storing sequence analysis results obtained from the project's analysis workflows. Data stored in these databases were obtained from various bioinformatics tools implemented as part of the protein domain recognition pipeline (see section 3.3.1), extracytoplasmic protein prediction pipeline (see section 3.3.2) and the protein similarity searches (see section 3.3.4). Each database was designed to store the output for a specific tool. These databases provide links between sequence information or phenotypic profiles and various protein sequence features. For example, number of transmembrane proteins across the taxa known to live in soil can be summarised using the information in the TMHMM results database. In the TMHMM result database, every protein with TMHMM prediction results are provided with their corresponding taxa of origin. The taxa information can then be used to link the TMHMM prediction results with another database that contains habitat information of the taxa. As a result, a mapping between microorganisms' habitats and the transmembrane sequence features can be made.

A database of predicted extracytoplasmic proteins

The predicted extracytoplasmic protein database (PredExtDB) is a structured relational database developed to store information about proteins with positive targeting and cell-surface anchoring feature prediction. This integrated database stores results from various prediction tools in the extracytoplasmic identification pipeline (see Section 3.3.2). The strategy used to include proteins into the PredExtDB is described later in section 3.3.3. The PredExtDB contains data of the predicted extracytoplasmic proteins. Each result includes a protein accession, the name of the analysis tool yielding the positive prediction and additional information about the result such as cleavage site, and topology. Therefore, PredExtDB is a collection of candidate putative extracytoplasmic proteins including transmembrane proteins, surface-associated proteins and secreted proteins (see Figure 2.8).

Instead of multiple queries, one per each result database of the prediction tools used, PredExtDB allows a single query to return results from all prediction tools. Thus, PredExtDB reduces the time needed to query a large and heterogeneous set of data of the protein subcellular localisation prediction results.

3.3 The sequence analysis pipelines developed in this project

Several bioinformatics tools were required in this project to provide information about protein sequence features. The sequence features of interest in this study included targeting signals, cell-surface anchors, functional regions. In this project, the sequence analysis pipelines were developed using Microbase as a framework. The pipelines chain together different bioinformatics tools and allow the analysis the be done in an appropriate order. A pipeline contains a set of relevant tools. Using Microbase enables several analysis pipelines to be processed in parallel.

A set of project-specific Microbase responders were developed in Java^{TM} in order to perform the various analysis tasks required by this project. The main protein sequence analysis pipelines constructed and implement in this project are: 1) a protein domain recognition pipeline; 2) extracytoplasmic protein prediction and filtering pipelines; 3) a protein similarity search pipeline. These pipelines consist of a number of responders that wrap several bioinformatics tools, including InterProScan, HMMER, SignalP, TMHMM, LipoP and BLAST. Computational work required by each of these tools proceeds in parallel. In this project, all bioinformatics-tools required input files of FASTA-formatted protein sequence data. Tool-specific analysis results generated by each responder were parsed into individual relational databases. Responders included in the analysis pipelines are described in detail in the following subsections.

3.3.1 Protein domain recognition pipeline

Existing bioinformatic tools were employed to identify characterised protein signatures. InterProScan was utilised to recognise well-characterised signatures in an integrated protein signature databases, while newly characterised project-specific protein domains (see Section 7) were detected using HMMER.

The InterProScanProcessor responder

The InterProScanProcessor responder wraps InterProScan version 4.4. This program identifies any known protein signatures including protein domains, motifs, families, functional sites, and GO term annotations on a given protein sequence. An InterProScan process could take a relatively long time. Running 100 protein sequences on InterproScan with all of the InterPro member databases (see Section 2.8.2) took approximately 1 hour on a typical desktop computer with 2 CPUs, and 2 GB of memory. In order to analyse several million protein sequences stored in the GPDB (see Section 3.4.1) within a reasonable amount of time, a restricted set of InterPro databases mainly associated with the characterisation of protein functions were used. As a result, InterProScan was configured to implement algorithms to search for those protein signatures stored in the following databases: Pfam database , SUPERFAMILY database , SMART and PROSITE databases. Options to annotate the results with GO terms and to annotate the results with the corresponding InterPro entries were switched on. The command line used was: `iprscan -cli -format raw -altjobs -iprlookup -goterms -appl hmmpfam -appl hmmsmart -appl superfamily -appl pattern-scan -appl profilescan -appl seg -i filename`

The output files produced by InterProScan were parsed and then stored in a custom database named 'InterPro result' database (see Figure 3.2).

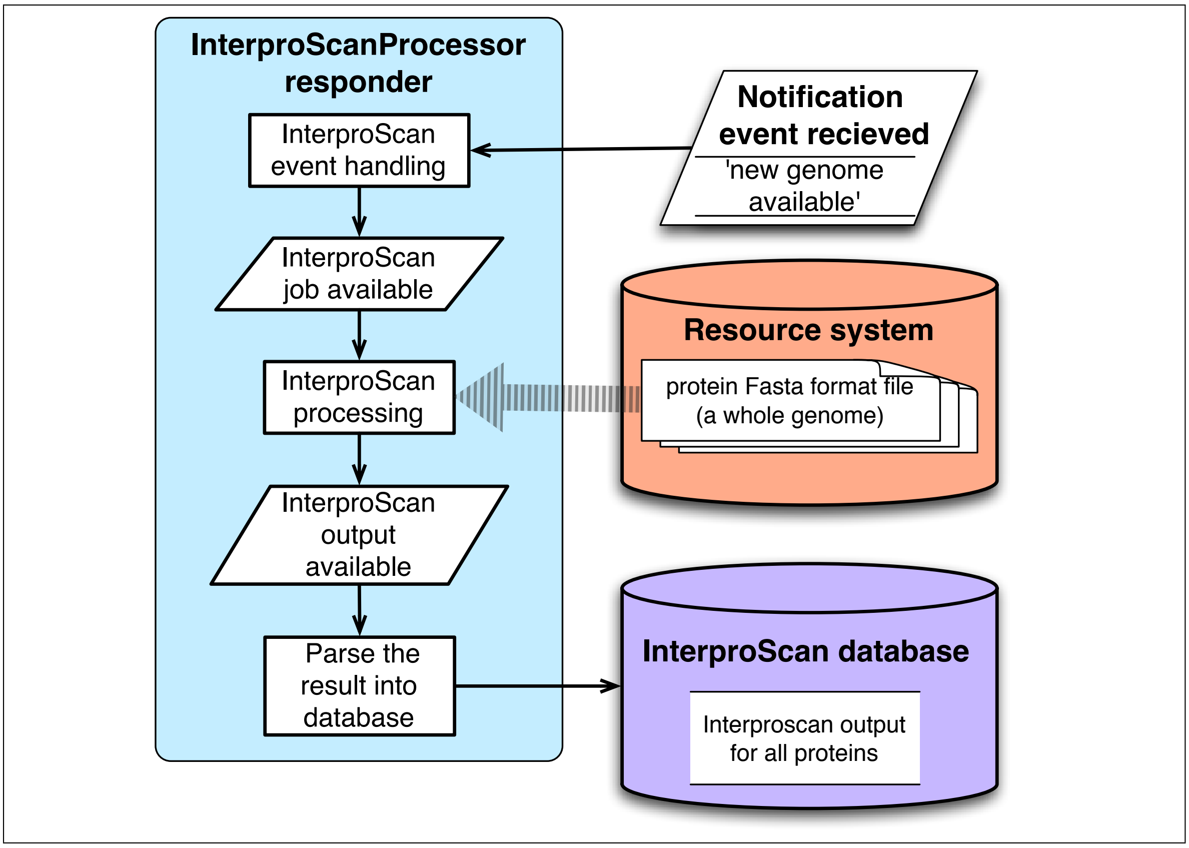

InterproScanProcessor responder
InterproScan event handling
InterproScan job available
InterproScan processing
InterproScan output available
Parse the result into database
Notification event recieved
'new genome available'
Resource system
protein Fasta format file (a whole genome)
InterproScan database
Interproscan output for all proteins

Figure 3.2: A flowchart diagram illustrating the operation of the InterProScanProcessor responder. The InterProScanProcessor responder is initiated by a 'new genome available' notification event. As a result, the InterProScan event handler splits the new proteome into blocks of around 100 proteins. Each block forms an InterProScan compute job entry. The jobs are then assigned to available worker computers, resulting in requests for associated input files (FASTA format) from the Microbase resource system. Once a computer receives its requested input files, the machine then executes InterProScan on the input protein sequences. On completion, the InterProScan output file is parsed and stored into the InterProScan result database.

The HmmerSearch responder

The HmmerSearch responder wraps HMMER 2.3.2. This responder is in charge of searching for a given HMM protein profile in all protein sequence data stored in the GPDB. In this project, a potentially novel mucosa-associated protein domain, termed M60-like domain was identified (discussed in Chapter 7). The HMM profile for M60-like was used by the HmmerSearch responder to search for the domain on all protein sequences in the GenomePool. The `hmmsearch` command was used with the default setting. The inclusion e-value for the default setting was 10. This responder can be used to search any new domain HMM profile that is not yet available in public databases.

3.3.2 The extracytoplasmic protein prediction pipeline

One of the goals of this project was to identify the putative microbial extracellular proteins of different microorganism's cell surface organisations by using existing bioinformatics tools. In order to make a universal prediction of extracytoplasmic proteins from primary amino acid sequences, several protein subcellular localisation prediction algorithms and tools were employed to detect well-characterised targeting signals and potential extracellular protein domains. These prediction tools include SignalP, LipoP, and TMHMM. SignalP is a widely used prediction tool designed to predict the N-terminal signal that targets precursor protein sequences to the Sec-pathway of both bacteria and eukaryotes. LipoP predicts prokaryotic lipoproteins which are anchored on the cell surface. TMHMM identifies alpha-helical transmembrane regions that allow proteins to be located through a lipid bilayer cytoplasmic membrane present in all cellular life forms. In addition, InterProScan was also employed to identify known extracellular or surface-associated protein signatures such as LPXTG or sortase motifs, porins and outer membrane signatures. The workflow was designed to provide an appropriate decision route best suited to sequences based on the cellular structures of their source organisms. For example, LipoP detects the presence of N-terminal signal peptidase II (SPII) cleavage site to identify putative prokaryotic lipoproteins. However, LipoP does not provide meaningful results for eukaryotic proteins so there is no point in processing eukaryotic proteins through LipoP.

In addition to executing bioinformatics tools, the workflow also takes into account GO terms referring to known surface and secreted protein domains for consideration as potential extracytoplasmic proteins. BLASTP searches were also employed as a strategy to identify sequences homologous to experimentally verified extracytoplasmic proteins to provide evidence that these proteins might be targeted the same subcellular localisation. The set of experimentally-verified extracytoplasmic proteins was derived from ePSORTdb v.2.0. Together, all the approaches used by the pipeline cover a wide range of strategies to predict whether a given protein sequence is potentially secreted or exposed to the surface of a cell. Each prediction tool was wrapped in a Microbase responder, allowing the construction of an automated sequence analysis workflow. Implementations of the responders developed for this pipeline are now described in detail. The application of these responders is discussed in Chapter 4.

It is notable that this study does not cover GPI anchors. A GPI anchor is a cell membrane anchoring structure found in some surface proteins of most eukaryotes [Omaetxebarria *et al.*, 2007]. Due to several practical reasons, the pipeline does not include any software to detect GPI-anchored proteins. First of all, GPI anchors are challenging to detect by both experimentally and computationally [Eisenhaber *et al.*, 1999][Eisenhaber *et al.*, 2000]. Currently available GPI anchor prediction software are organism specific and are therefore not reliable for all eukaryotes due to a limit set of experimental data for a training purpose [Eisenhaber *et al.*, 1999]. Secondly, not every eukaryote included in this study is known to possess GPI-anchored proteins e.g. *T. vaginalis* [Hirt *et al.*, 2007]. Lastly, GPI-anchored proteins typically have N-terminal signal peptides and C-terminal hydrophobic regions [Howell *et al.*, 1994][Eisenhaber *et al.*, 1999]. Therefore, these proteins can be indirectly detected by SignalP and TMHMM included in the pipeline developed in this project.

The extracytoplasmic protein prediction pipeline is composed of a set of 'processing' responders and a 'filtering' responder. The processor responders execute different protein subcellular localisation prediction tools and then parse the analysis result into structured databases. The result-filtering responder was developed for the purpose of extracting all positively predicted protein sequences identified by one or more responders.

The SignalPProcessor responder

The SignalPProcessor responder is responsible for executing the SignalP tool in order to predict the presence of N-terminal signal peptides and their corresponding cleavage sites in protein sequences. SignalP version 3.0 [Dyrlovbendtsen, 2004] was used by this responder.

Only the first 70 amino acid residues at the N-terminus of protein sequences were used as an input to SignalP as recommended by the tool developers [10]. The length limit was suggested based on the finding that an N-terminal signal peptide is seldom longer than 45 amino acids [Nakai, 2000]. SignalP was configured to use both HMM and NN prediction algorithms. SignalP provide an option

[10] http://www.cbs.dtu.dk/services/SignalP/instructions.php, accessed accessed 20th October 2010

to select the specific organism type being analysed (Gram+/Gram-/Euk). This selection results in the use of an appropriate training data set for the tool's algorithms. The SignalPProcessor responder used for executing SignalP within the automated pipeline selects the appropriate SignalP options based upon organisms taxonomy information stored in the GPDB (see Figure 3.3).

SignalP 3.0 was trained by the SignalP developers on sets of amino acid sequences from Gram-positive and Gram-negative prokaryotes, as well as eukaryotes. However, many of the sequences stored in the GPDB are encoded by several other organism groups including archaea, divergent eukaryotes, and known non Gram-staining bacteria such as Mycoplasma. Therefore, to process all sequences from the GenomePool through the SignalPProcessor responder, information regarding an organism's type retrieved from the GPDB before SignalP is executed. This information is provided by the the NCBI taxonomic lineage of the organism encoding the set of protein sequences to be analysed (the value of the `ORGANISM` tag from the original GenBank file).

The first part of an organisms' taxonomic annotation normally denotes its superkingdom level. This information is checked by the SignalPProcessor responder to identify the source organism of a protein sequence as either a prokaryote or eukaryote. This inspection determines whether SignalP is executed with the option `euk`, in case where the source organism is an eukaryote. On the other hand, if a prokaryote is detected, SignalP is instructed to execute twice: once with the option `gram-`, and the second time with the option `gram+`. Two runs were applied to all prokaryotic sequences including sequences from archaea, Gram-staining bacteria and other non Gram-staining bacteria. At first glance, running SignalP twice seems unnecessarily wasteful of compute resources. However, as the run time for SignalP is not hugely computationally intensive, it was decided to simplify the design of the responder by not taking account of the Gram-stain type. Moreover, for non Gram-staining bacteria, there is no appropriate command line option on SignalP. Executing SignalP twice on a set of sequences provides a greater selections of prediction results for non Gram-staining prokaryotic protein sequences. SignalP has not been trained for use with non Gram-staining bacteria, and therefore the prediction results of these proteins may not be as accurate as the results of Gram-staining bacteria. However, these less accurate results may still be useful when integrated with evidence provided by the other analysis tools used within this project.

SignalP provides a number of options regarding the data formatting of it output. The 'short' output format was used for this work. All fields from the output were parsed from the generated output files, and stored in a relational database. The prediction results from SignalP were utilised by a result-filter responder in a later workflow step (see Section 3.3.3).

The TMHMMProcessor responder

The TMHMM algorithm is employed to predict alpha helical transmembrane regions on protein sequences; TMHMM version 2.0 [Krogh *et al.*, 2001] was used for this project. The TMHMMProcessor responder executes the TMHMM algorithm.

All protein sequences in the GenomePool from all three domains of cellular life were processed by this responder (see Figure 3.4). The prediction results generated in the 'short' output format were parsed and inserted into a structured database.

The LipoPProcessor responder

The LipoPProcessor responder is responsible for executing the LipoP on appropriate protein sequences (see Figure 3.5). LipoP is designed to predict N-terminal lipoprotein signal peptide cleavage sites. LipoP was trained with a Gram-negative bacterial data set, but can also detect Gram- positive lipoproteins. In this work, LipoP was used to analyse prokaryote proteins. The tool was also applied to archaeal proteins, since lipoproteins can also be found in archaea [Eichler and Adams, 2005]. LipoP version 1.0 was used for the analyses carried out in this work [Juncker *et al.*, 2003].

The developers of LipoP recommend that only the first 60 N-terminus amino acids from protein sequences are used as input to the tool in order to prevent erroneous events. The superkingdom of an organism from which a protein derived was checked using the organism's taxonomic lineage as stored in the GPDB (see Section 3.3.2 for more details). All prediction results were generated in the 'short' output format and then parsed into a database. The 'short' output format a tabular format that is straightforward to parse.

The SCLBlastPProcessor responder

Another technique that may be used to identify potential extracytoplasmic protein candidates is the inference through sequence similarity. If a primary protein sequence was homologous to a known extracytoplasmic sequences, then there is a possibility that the two sequences might have the same localisation site [Gardy, 2004].

The SCLBlastPProcessor responder employs BLAST (version 2.2.19) to query input sequences against sets of proteins with known subcellular-localisations. BLASTP is used to search every protein sequence stored in the GPDB against a database of experimentally verified bacterial outer membrane and extracellular protein sequences were derived from the ePSORT database (see Figure

3.6)(see Background section 2.8.1 for more details). The set of experimentally verified sequences include proteins classified in ePSORTdb (v.2.0) as being ‘Cellwall’, ‘Extracellular’, ‘Outer membrane’, ‘Periplasmic/outermembrane’, or ‘CytoplasmicMembrane/Cellwall’.

A cut-off e-value of 1×10^{-4} was used for running the BLASTP program in the pipeline. The search results were obtained in the ‘m8’ format; a tabular format that is straightforward to parse. All fields in the BLAST output file were parsed and stored in a relational database for further analyses.

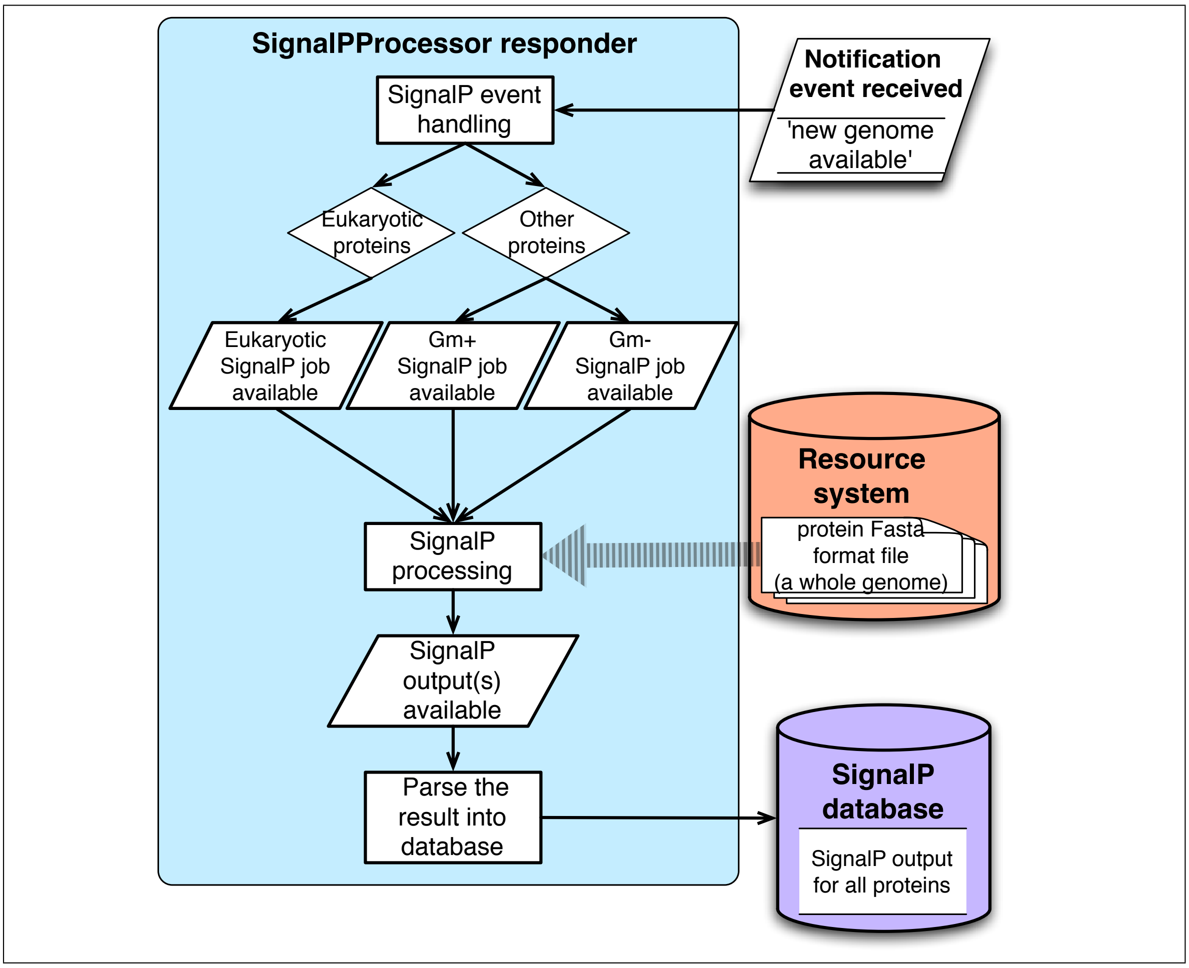

SignalPProcessor responder
SignalP event handling
Notification event received
'new genome available'
Eukaryotic proteins
Other proteins
Eukaryotic SignalP job available
Gm+ SignalP job available
Gm- SignalP job available
Resource system
protein Fasta format file (a whole genome)
SignalP processing
SignalP output(s) available
Parse the result into database
SignalP database
SignalP output for all proteins

Figure 3.3: A flowchart diagram illustrating the operation of the SignalPProcessor responder. The SignalPProcessor responder is initiated by a 'new genome available' notification event. As a result, the SignalP event handler component produces a list of SignalP jobs for the new fragment of protein sequences. The SignalP event handler considers primary protein sequences from prokaryotes and eukaryotes differently. Given a set of eukaryotic proteins, a SignalP job with the appropriate eukaryotic command line switch was generated. In the case of non-eukaryotic (bacterial and archaeal) protein sequences, two separate SignalP jobs were generted: one with the Gram-positive command line option, and the other with the Gram-negative options set. Jobs are assigned by Microbase to an available worker machines. The computers request the appropriate input files (FASTA format) from the Microbase resource system. Once request input files are retrieved, the machines then executes the SignalP software with the assigned command line options for the input protein sequences. Once the processing has completed, the SignalP output files are parsed and stored in the SignalP result database.

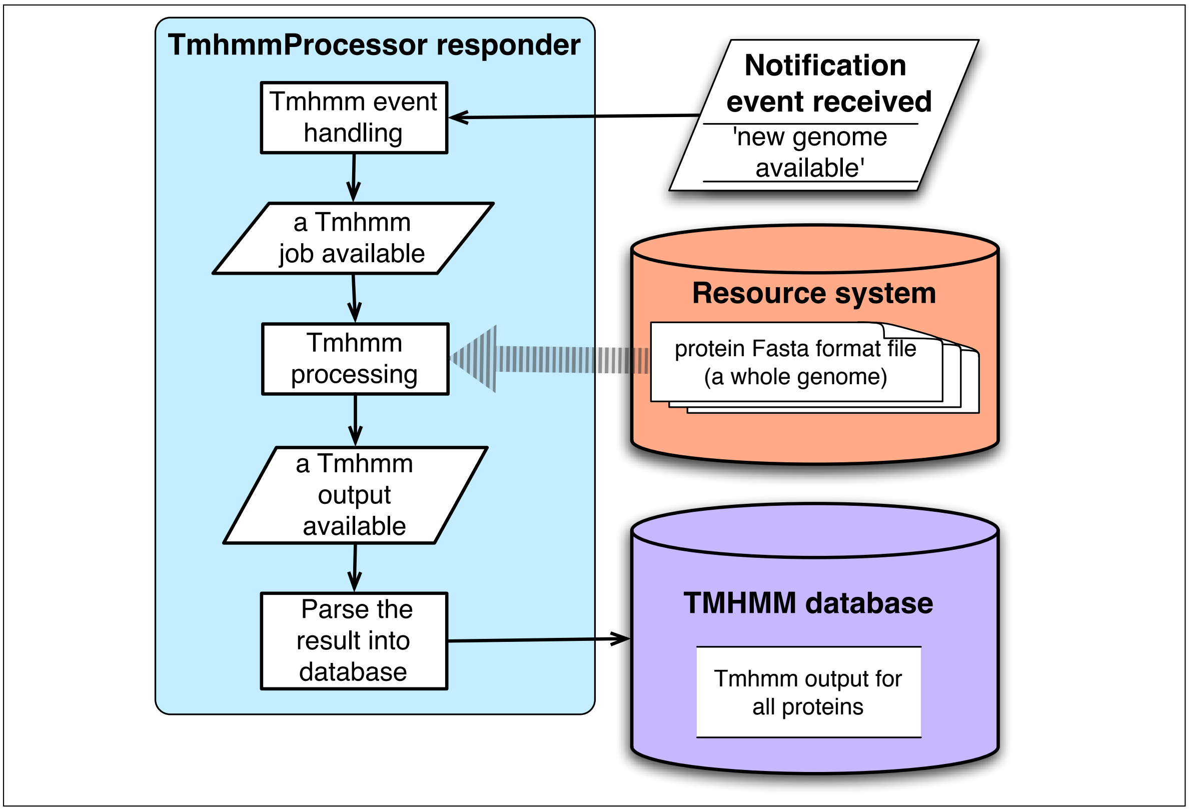


Figure 3.4: A flowchart diagram illustrating the operation of the TMHMMProcessor responder. The TMHMMProcessor responder is initiated by a 'new genome available' notification event. As a result, the TMHMM event handler component produces a TMHMM job to analyse the protein sequences encoded by each genome fragment. TMHMM jobs are then assigned to available computers by Microbase. Necessary input files (FASTA format) are requested from the Microbase resource system. Worker machines then execute the TMHMM algorithm on the input protein sequences. When processing has completed, the generated TMHMM output files are parsed and stored in the TMHMM result database.

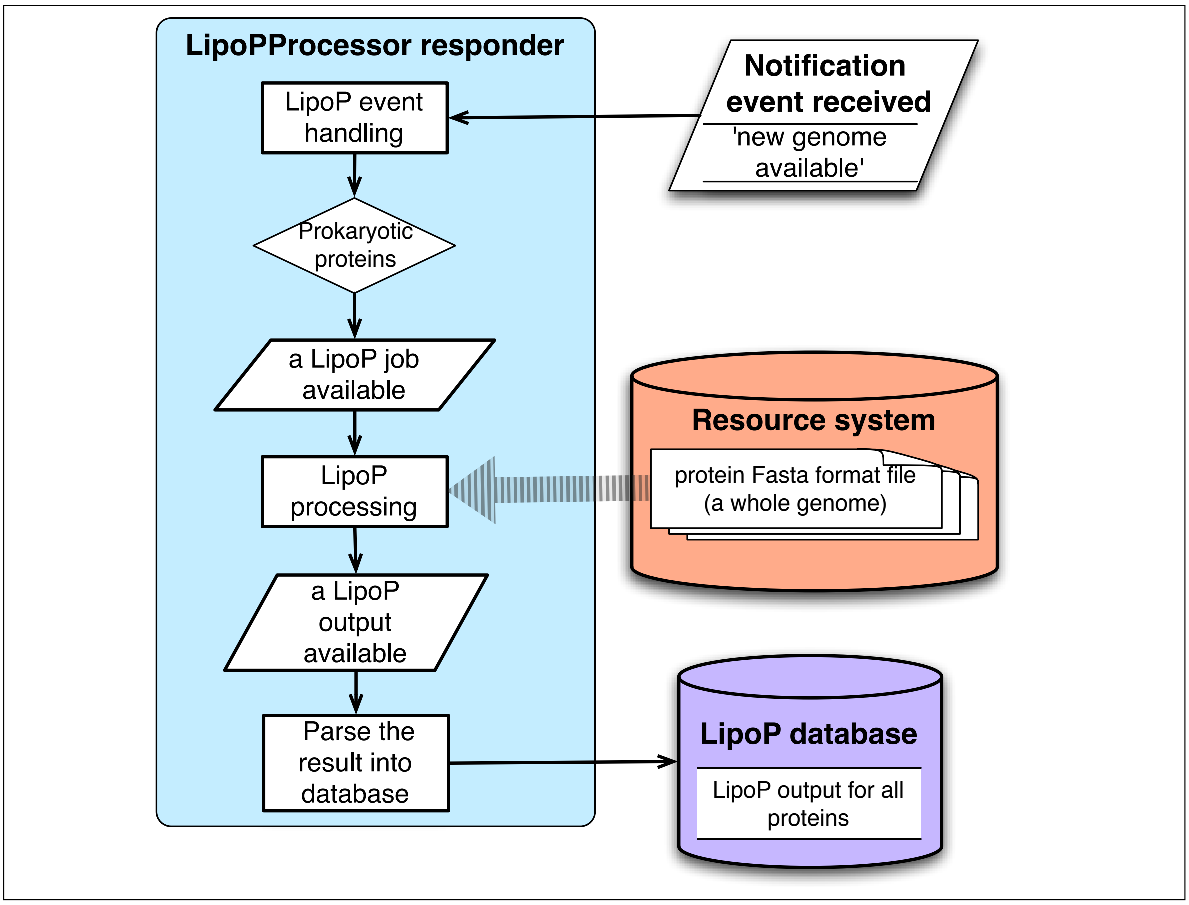


Figure 3.5: A flowchart diagram illustrating the operation of the LipoPProcessor responder. The LipoPProcessor responder is initiated by a 'new genome available' notification event. The LipoP event handler component only generates Microbase jobs for prokaryotic proteins since eukaryotes do not have N-terminal lipid anchors. LipoP was only designed to identify prokaryotic lipoproteins. As a result, the LipoP event handler produces a LipoP job for each proteome encoded by a given prokaryotic genome. These jobs are then assigned to an available worker computers, resulting the necessary input files (FASTA format) being transferred from the Microbase resource system. The worker machines then execute the LipoP software on the input protein sequences. On process completion, the LipoP output files are parsed and stored in the LipoP result database.

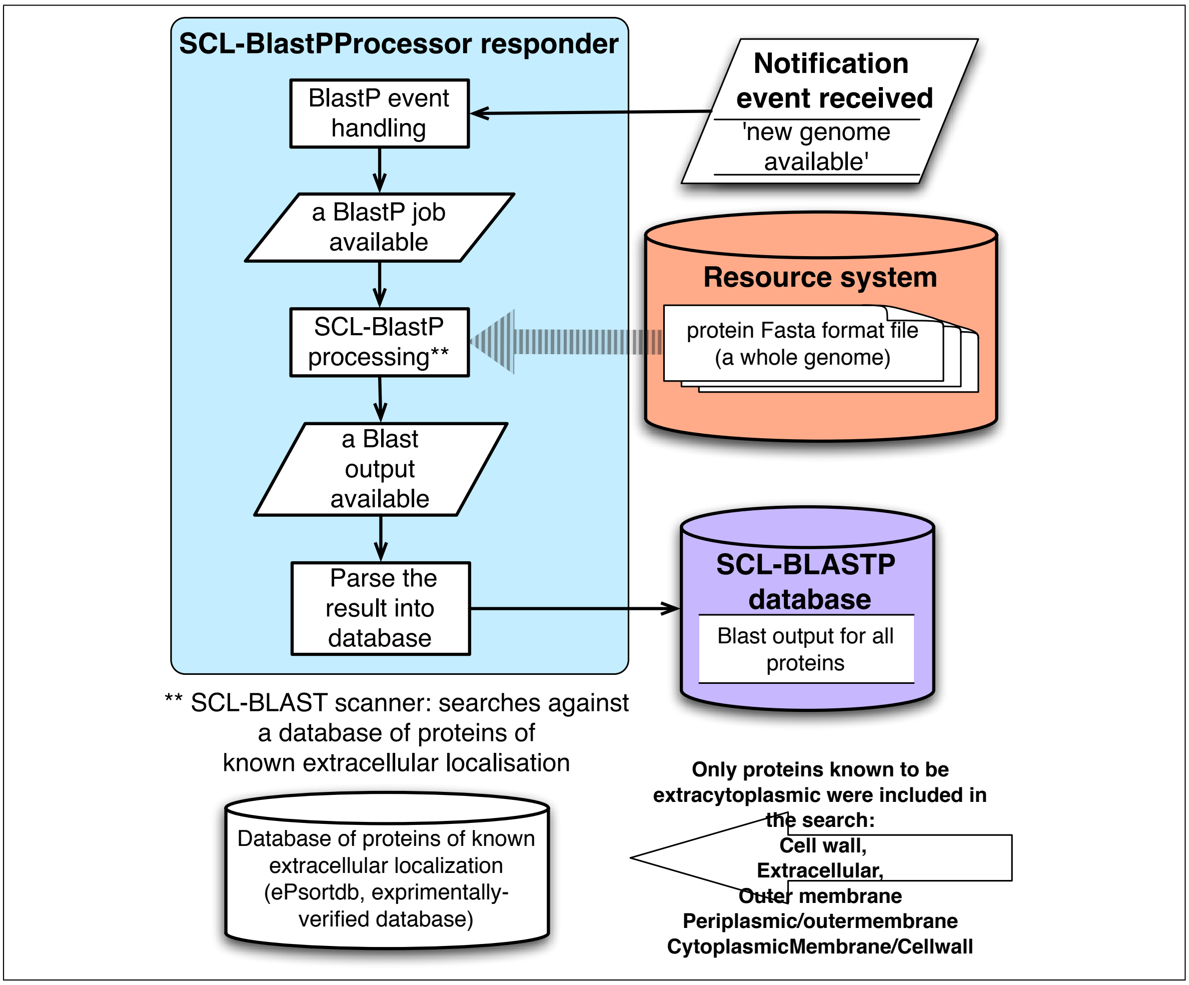

SCL-BlastPProcessor responder
BlastP event handling
a BlastP job available
SCL-BlastP processing**
a Blast output available
Parse the result into database
Notification event received
'new genome available'
Resource system
protein Fasta format file (a whole genome)
SCL-BLASTP database
Blast output for all proteins
** SCL-BLAST scanner: searches against a database of proteins of known extracellular localisation
Database of proteins of known extracellular localization (ePsortdb, exprimentally-verified database)
Only proteins known to be extracytoplasmic were included in the search:
Cell wall,
Extracellular,
Outer membrane
Periplasmic/outermembrane
CytoplasmicMembrane/Cellwall

Figure 3.6: A flowchart diagram illustrating the operation of the SCLBlastPProcessor responder. The SCLBlastPProcessor responder is initiated by a 'new genome available' notification event. The BLASTP event handler component produces a SCL-BLASTP job for all the protein sequences. The job is then assigned to an available worker machine. The appropriate input files (FASTA format) are requested from the Microbase resource system, and the machines then execute the BLASTP algorithm on the input protein sequences. The input sequences were queried against a set of experimentally verified extracytoplasmic proteins obtained from ePSORTdb version 2.0. Extracytoplasmic proteins include proteins that have been verified to be located on the cell wall, the outer membrane, cytoplasmic membrane, periplasmic space, or are extracellular. When BLASTP works is completed, the output files are parsed and stored in the SCL-BLASTP result database.

3.3.3 The extracytoplasmic proteome filtering responder

The responder described in the previous sections are responsible for executing variety of protein sequence analysis tools. As a results, a set of independent result databases are populated that contain analysis results for proteins stored in the GPDB. Not every protein is of interest to this study. The extracytoplasmic proteome filtering responder is responsible for extracting all positive results produced by the various bioinformatics tools described in the previous section. For a given protein, the results from each prediction tool were assessed to determine whether the protein should be included in the 'Predicted Extracytoplasmic Protein Database' (PredExtDB) (see Section 3.2.4). A summary of the targeting features from each analysis tool for each putative extracytoplasmic protein is stored in the PredExtDB (see Figure 3.7).

The extracytoplasmic proteome filtering responder was designed to process protein localisation predictions from archaea, bacteria and microbial eukaryotes for the purpose of filtering out cytoplasmic protein sequences, or sequences without any cell surface targeting signals. Prediction results from the bioinformatics tools incorporated into the workflows described in the previous sections (Section 3.3.2, 3.3.1) were processed by this responder. Appropriate localisation evidence was selected for each protein, based on the taxonomic group of the source organism and their cell surface structures. For example, a protein from a Gram-positive organism might be assessed based on the result from the SignalP program executed with option `gram+`.

The filtering responder is initiated by a user-generated notification event. Once triggered, a list of genome fragment accessions is read from the GPDB. Each accession number is used as an input query to extract associated information from the various prediction result (secondary) databases generated by the extracytoplasmic protein prediction pipeline 3.3.2), and the InterProScanProcessor responder (see Section 3.3.1). To determine whether a particular protein is included in the PredExtDB, a number of tool-specific filtering strategies are employed. The result filtering strategy for each tool and the implications for different organism types (eukaryote, archaea, Gram-negative and Gram-positive bacteria) are described below.

- SignalP result filtering: This filtering step is responsible for recruiting all candidate exported proteins into the PredExtDB. Firstly, the accession number is checked against a local copy of the NCBI taxonomic database to initially identify its organism superkingdom as a prokaryote or eukaryote. If the genome is derived from prokaryote, further investigation is performed to determine whether it is a Gram-positive or Gram-negative bacteria, or neither. The information of bacterial gram staining was checked against a locally-installed copy of the GOLD genome

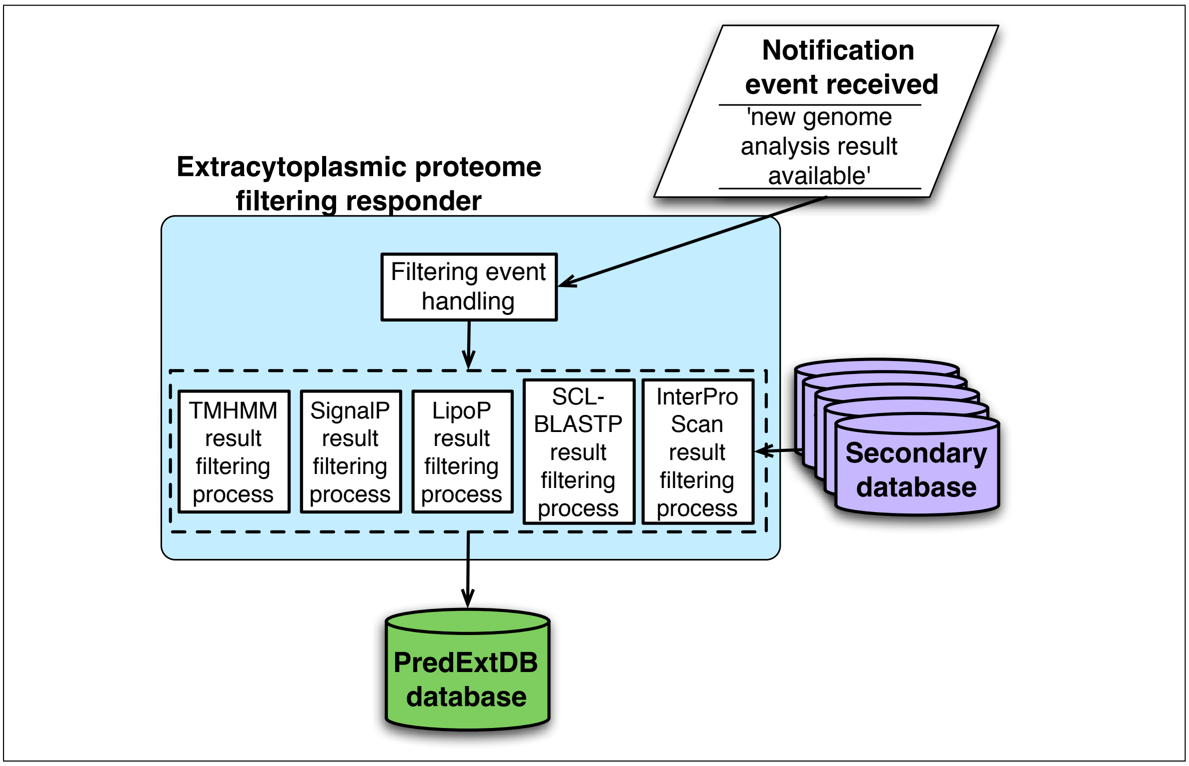

Notification event received
'new genome analysis result available'
Extracytoplasmic proteome filtering responder
Filtering event handling
TMHMM result filtering process
SignalP result filtering process
LipoP result filtering process
SCL-BLASTP result filtering process
InterPro Scan result filtering process
Secondary database
PredExtDB database

Figure 3.7: A flowchart diagram illustrating the operation of the Extracytoplasmic proteome filtering responder. All sequence analysis results were filtered in order to include the proteins that are likely to be extracytoplasmic proteins into the PredExtDB for further analysis.

information database. For a given genome accession number, the appropriate SignalP results are extracted from the secondary database containing SignalP prediction results. The extracted results are then checked for a positive prediction. SignalP results were considered ‘positive’ if an N-terminal signal peptide was predicted to be present by either the NN or HMM algorithms (Figure 3.8).

- TMHMM result filtering: The purpose of the TMHMM result filtering step is to select proteins that are likely to be transmembrane-located and store them into the PredExtDB. Given a genome accession number, the associated TMHMM prediction results are extracted from the TMHMM job executions. A result for a particular protein is considered as ‘positive’ if at least one alpha-helix region was predicted. An overview of the data flow for this process is shown in Figure 3.9.

- LipoP result filtering: The LipoP result filtering step is responsible for filtering candidate microbial lipoproteins or secreted proteins into the PredExtDB. A protein is considered as a putative lipoprotein if the protein was predicted by LipoP to have an N-terminal cleavage site that is recognised by SPII. The LipoP prediction results were obtained from the LipoP result database (see Figure 3.10).

- SCL-BLASTP result filtering: The SCL-BLASTP filtering step was designed to identify sequences that are highly similar and by extension, potentially homologous to experimentally verified extracytoplasmic proteins obtained from the ePSORTdb. The inclusion criteria used were an e-value cutoff of 1×10^{-9} and a requirement that the length of both query and subject sequences must range from 80-120% of each other. An overview of the data flow is shown in Figure 3.11

- Surface-associated protein domain and GO term result filtering: a protein sequence is filtered if the sequence was annotated to own at least one of the sequence features listed in Table 3.1. Protein domains were used to referred to GO terms.

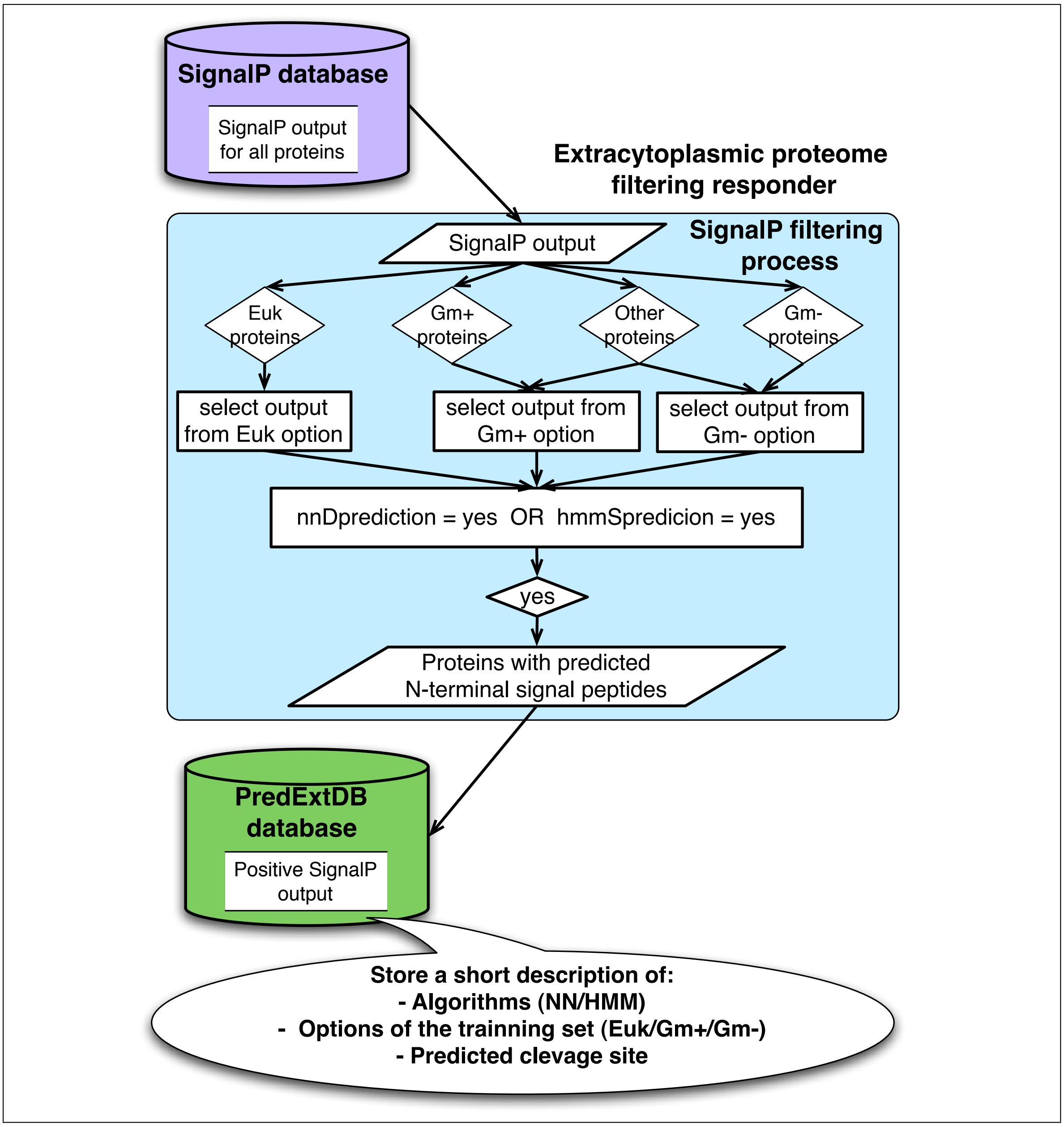

SignalP database
SignalP output for all proteins
Extracytoplasmic proteome filtering responder
SignalP filtering process
SignalP output
Euk proteins
Gm+ proteins
Other proteins
Gm- proteins
select output from Euk option
select output from Gm+ option
select output from Gm- option
nnDprediction = yes OR hmmSpredicion = yes
yes
Proteins with predicted N-terminal signal peptides
PredExtDB database
Positive SignalP output
Store a short description of:
- Algorithms (NN/HMM)
- Options of the trainning set (Euk/Gm+/Gm-)
- Predicted clevage site

Figure 3.8: A flowchart diagram illustrating the operation of the SignalP result filtering responder. Depending on the taxonomic group an organism belongs to, its proteins may have been analysed by SignalP either one or two times. The SignalP result filtering process was designed to iterate over the entire collection of protein sequences and query the appropriate SignalP results from the SignalP database. If a protein sequence came from an eukaryote, Gm+ or Gm-, then the SignalP database was queried for results generated with the `euk`, `gram+`, `gram-` command line options, respectively. If a protein sequence was from any other taxonomic group, then the SignalP database was queried for results generated with the `gram+` and `gram-` settings. A protein was considered to have a positively-predicted N-terminal signal peptide, if either the NN or HMM algorithms reported a positive prediction. The 'nnDprediction' and 'hmmSprediction' fields were were used as inclusion criteria for the NN and HMM algorithms, respectively. 'nnDprediction' is a summarised NN-prediction result based on a score used as the criteria for discrimination of secretory and non-secretory proteins. 'hmmSprediction' is the prediction result based on the probability score of a signal anchor calculated using HMM. The positive SignalP results that met these criteria were parsed into the PredExtDB with their predicted cleavage site, algorithm and training dataset option used.

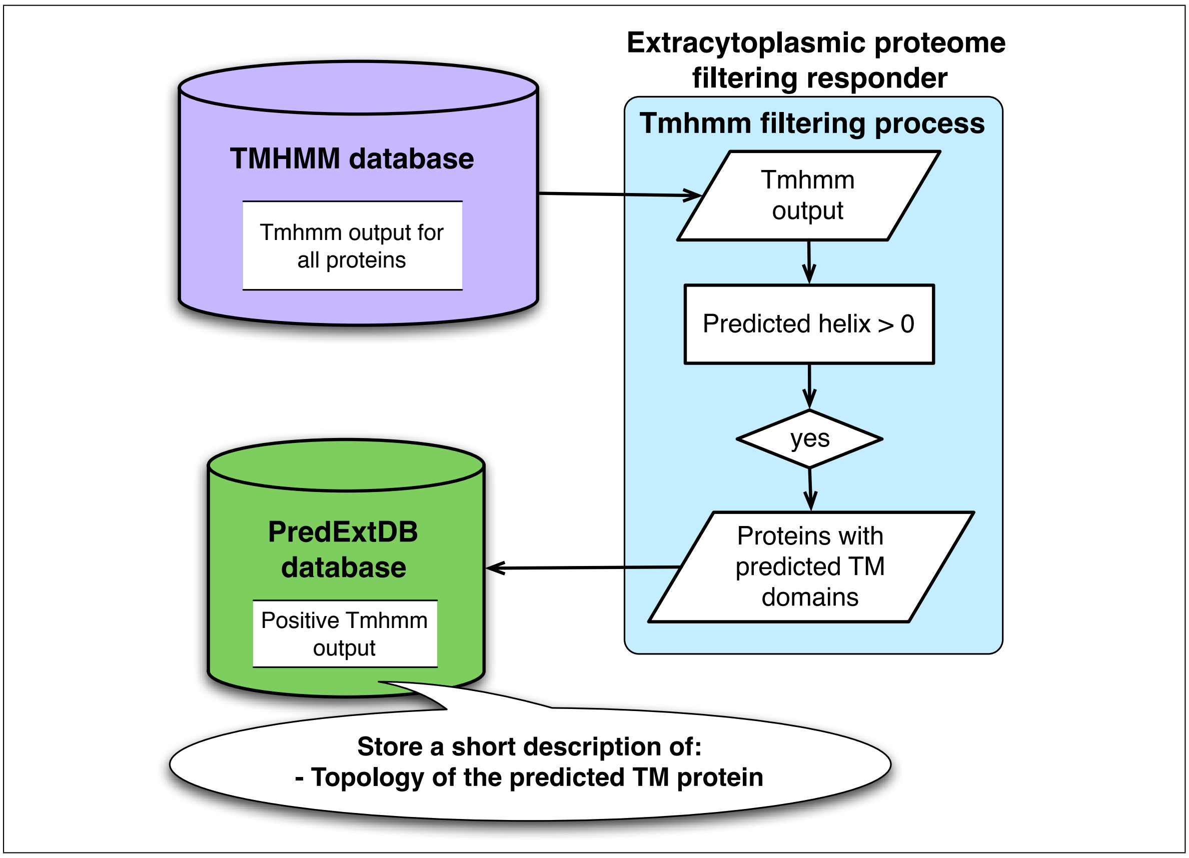

Extracytoplasmic proteome filtering responder
Tmhmm filtering process
TMHMM database
Tmhmm output for all proteins
Tmhmm output
Predicted helix > 0
yes
Proteins with predicted TM domains
PredExtDB database
Positive Tmhmm output
Store a short description of:
- Topology of the predicted TM protein

Figure 3.9: A flowchart diagram illustrating the operation of the TMHMM result filtering process. All TMHMM results were filtered in order to remove the proteins that are not likely to be transmembrane-located. TMHMM results with a positive 'predicted helix' value were copied into the PredExtDB for further analysis.

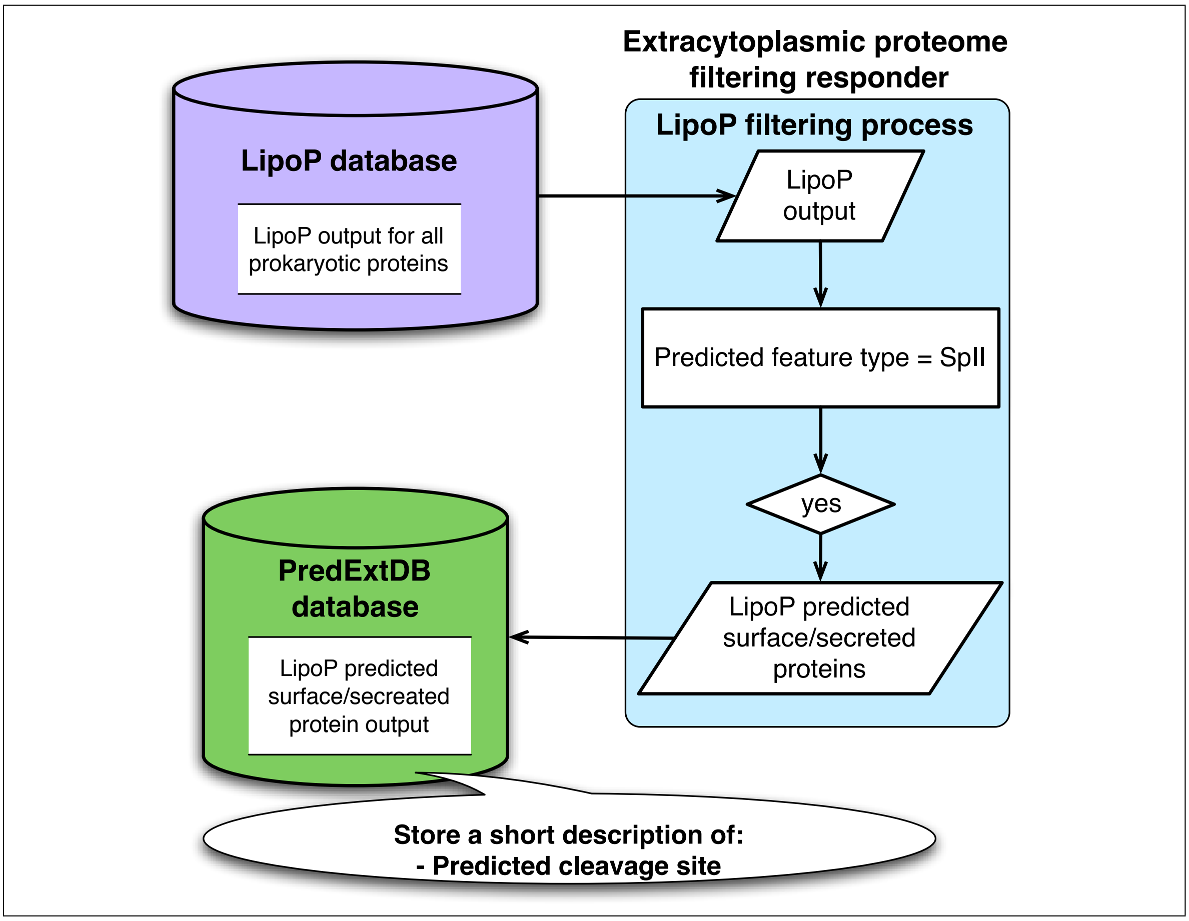

Extracytoplasmic proteome filtering responder
LipoP filtering process
LipoP database
LipoP output for all prokaryotic proteins
LipoP output
Predicted feature type = SpII
yes
LipoP predicted surface/secreted proteins
PredExtDB database
LipoP predicted surface/secreated protein output
Store a short description of:
- Predicted cleavage site

Figure 3.10: A flowchart diagram illustrating the operation of the LipoP result filter process. Filtering LipoP results involved copying predicted lipoproteins into the PredExtDB. Each prediction result from the LipoP database was examined. Proteins predicted to have features that could be recognised to be processed by signal peptidase II (SpII) were regarded as positive predictions.

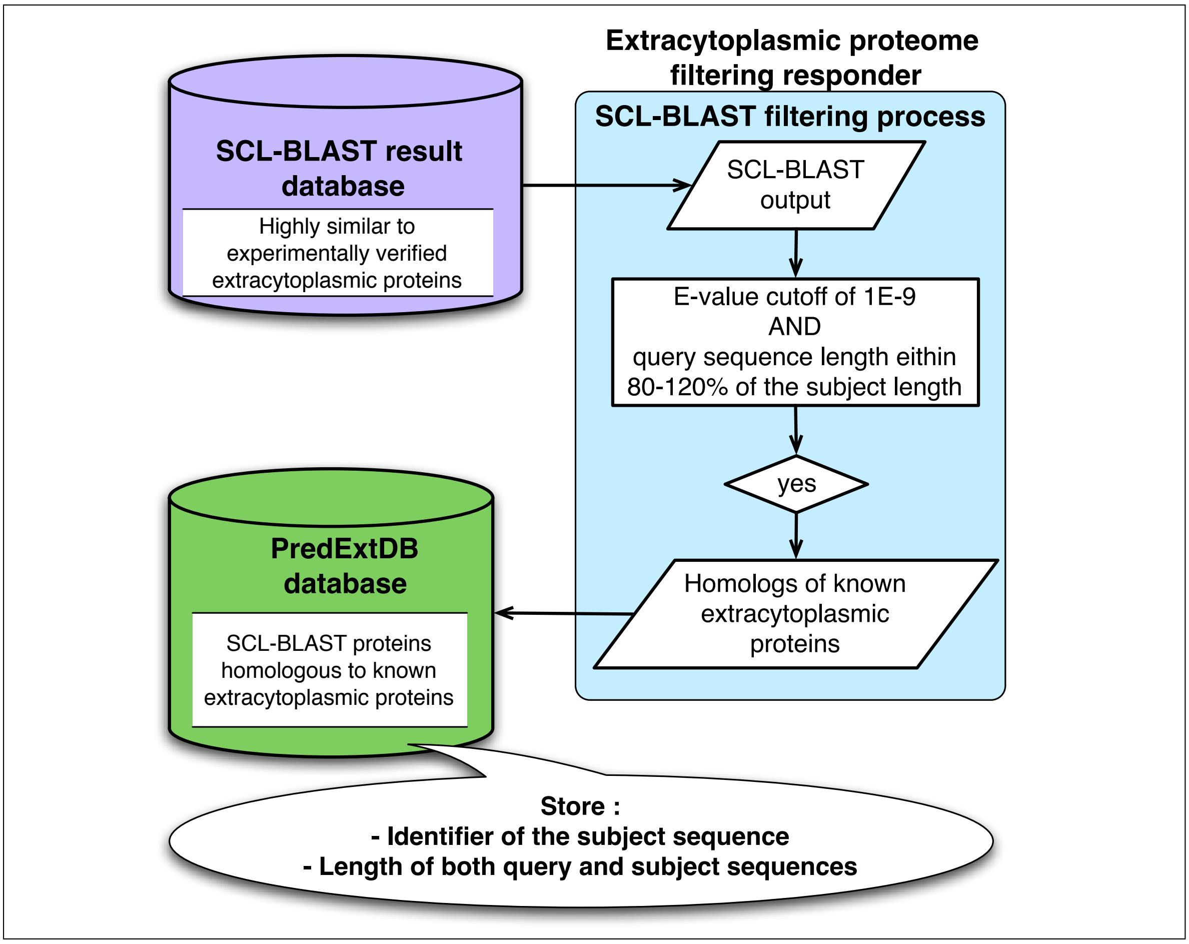

Extracytoplasmic proteome filtering responder
SCL-BLAST filtering process
SCL-BLAST result database
Highly similar to experimentally verified extracytoplasmic proteins
SCL-BLAST output
E-value cutoff of 1E-9 AND query sequence length eithin 80-120% of the subject length
yes
Homologs of known extracytoplasmic proteins
PredExtDB database
SCL-BLAST proteins homologous to known extracytoplasmic proteins
Store :
- Identifier of the subject sequence
- Length of both query and subject sequences

Figure 3.11: A flowchart diagram illustrating the operation of the SCL-BLAST result filter process. This result filtering process attempted to find only sequences that were highly similar to an experimentally-verified extracytoplasmic protein. Suitable proteins were then copied into the PredExtDB. A protein was considered as a potential homologs of a known extracytoplasmic proteins if its BLAST hit e-value was less than 1×10^{-9} and the length of the query sequence fell within 80-120% of the length of the subject sequence (known extracytoplasmic sequence).

Table 3.1: Protein motifs and Gene Ontology term known to presented on extracytoplasmic proteins. These protein signatures were used by the result filtering responder as criteria to consider a protein sequence as a putative extracellular protein. The protein motifs were obtained from various public databases containing protein signatures. The entry accessions represent accession numbers of particular protein signatures from the different databases. The first 2-3 alphabets in the accessions denote the database from which the signature was derived; PS = Prosite entry, PF = Pfam entry, SSF = Superfamily entry, GO = Gene ontology entry.

Feature description	**Entry accessions**
Gram-negative bacterial outer membrane/extracellular motifs	
General diffusion Gram-negative porins signature	PS00576
Enterobacterial virulence outer membrane protein signature	PS00694, PS00695
Fimbrial biogenesis outer membrane usher protein signature	PS01151
Bacterial type II secretion system protein D signature	PS00875
Aspartyl proteases, omptin family signature	PS00834, PS00835
OmpA-like domain	PS01068
Aerolysin type toxins signature	PS00274
Hemolysin-type calcium-binding region signature	PS00330
Gram-positive bacterial Cell wall/extracellular motifs	
S-layer homology	PS51272
Staphylococcal enterotoxin/Streptococcal pyrogenic exotoxin signature	PS00277, PS00278
Staphylocoagulase repeat signature	PS00429
Thermonuclease domain profile	PS50830
Other known protein features characterizing surface/secreted protein	
Bacterial extracellular solute-binding proteins	PS01037, PS01039, PS01040
Leucine rich repeat	PF00560
M protein motif	PF00746
Sortase motif	PF04203
LPXTG motif	PS50847
GW domain	SSF82057
NLPC P60	PF00877
LYSM protein	PF01476
Extracellular-related GO terms (cellular component)	
Extracecllular region/space	GO:0005576, GO:0044421, GO:0005615
Extracellular matrix	GO:0031012
Cell surface	GO:0009986
Cell wall	GO:0005618
Gram negative cell wall	GO:0009276
Outer membrane	GO:0009279, GO:0019867
Outer membrane periplasmic	GO:0030288
Extrinsic to membrane	GO:0019898

3.3.4 Responders for protein similarity search

The BLAST algorithm (version 2.2.19) was employed for sequence similarity searches. All protein sequences in the GPDB were subject to two separate searches. Two responders were developed in order to perform: 1) an all-vs-all search; 2) a search against all proteins from RefSeq database.

The all-against-all BLASTP responder

The all-vs-all BLASTP responder is responsible for performing bi-directional BLASTP search of all proteins in the GPDB against each other. Given the large number of protein sequences (approximately 3 million sequences), an immense amount of computational effort is required. It is not feasible to execute this analysis on a single computer [Gardner *et al.*, 2006][Shah *et al.*, 2007]. The responder splits the task into more manageable units of work. The pairwise comparison of each proteome is defined as an individual Microbase job. Once the computational work has been split into jobs, Microbase is responsible for overseeing the job execution. On completion of these jobs, result data is passed back to the BLAST responder. The data from each job is collated and inserted into a relational database for future use. The all-vs-all BLASTP responder is in charge of performing bi-directional BLASTP search of all proteins in the GPDB against each other. The all-vs-all BLASTP e-values generated by this responder were used at a later stage as sequence similarity scores to construct protein clusters (see Chapter 6, Section 6.2.2).

all-vs-RefSeq BLASTP responder

Similar to the all-vs-all BLASTP responder described above (see Section 3.3.4), the all-vs-Refseq BLASTP responder is responsible for splitting a large amount of computational work into jobs of a reasonable size for performing BLASTP searches on Microbase worker machines. This responder ensures that every protein stored in the GPDB is searched against the set of proteins in the NCBI RefSeq database. The pre-formatted BLAST database of the NCBI protein reference sequences was downloaded from the NCBI FTP site[11] (accessed 25 October 2009). The all-vs-RefSeq BLASTP results were used to identify protein sequences that are statistically associated with mucosal organisms (see Chapter 6, Section 6.2.2).

3.4 Results

This section describes statistical reports of the total genome sequence information in the GPDB and the performance of the sequence analysis workflows using the Microbase framework. A summary of the analysis results generated from each responder-encapsulated bioinformatics tools is also provided.

3.4.1 The GenomePool database

The GPDB contains 3,127 complete genome sequences and contigs from 867 microorganisms including archaea, bacteria and selected microbial eukaryotes. The total number of protein sequences

[11] ftp://ftp.ncbi.nih.gov/blast/db

Table 3.2: Summary of data incorporated in the GenomePool database. Number of organisms, genome sequences and protein-coding gene sequences were summarised in relation to the three domains of cellular life. The asterisk includes bacterial and archaeal chromosomes, as well as plasmids when present, and eukaryotic chromosomes and organelle genomes. In some cases, genome sequence data are spread over a number of contigs and/or scaffolds in so called draft genomes – in particular, for large eukaryotic genomes with important fractions of repetitive sequences.

Number of	Archaea	Bacteria	Eukaryote	Total
Microorganisms	55	780	32	867
Genome fragments*	80	1,556	1,491	3,127
Protein sequences	133,026	2,616,075	272,389	3,021,490

stored in the GPDB is 3,021,490. Bacterial protein sequences account for a major proportion (86.6%) of the sequences deposited in the GPDB. Out of the the 3 million sequences, roughly 23% are from the Gram-positive bacterial group (members of phyla Actinobacteria, Firmicutes and Tenericutes). 64% of the sequences belong to the Gram-negative bacterial group (members of bacterial phyla that are not in the Gram-positive group) (see Figure 3.12). Approximately 10% and 4% of the proteins are from microbial eukaryotes and archaea, respectively. Not surprisingly, the most common phyla in the GPDB are Proteobacteria and Firmicutes, respectively. Genome sequences of these two bacterial phyla are abundantly available since their members are predominantly found to be associated with humans as either pathogens or mutualists. There are approximately 2.6 times (421:159) more taxa from Proteobacteria than those from Firmicutes. These two phyla account for 66.8% of the genomes in the GPDB. 3.6% are from microbial eukaryotes (protist and fungi) while 6.3% are archaeal genomes. The number and distribution of genomes in the GPDB are shown in Table 3.2 and Figure 3.12, respectively. Other bacterial phyla include Actinobacteria, Tenericutes, Acidobacteria, Aquificae, Bacteroidetes, Chlamydiae, spirochaetes, Thermi, Verrucomicrobia, Dictyglomi, Elusimicrobia, Fusobacteria, Nitrospirae and Planctomycetes. Archaeal genomes include 36 taxa from Euyarchaeota, 17 taxa from Crenarchaeota, and one of each of Korarchaeota and Nanoarchaeota. Genomes from microbial eukaryotes included in the GPDB covered Fungi and Protists. The Fungi genomes sequences comprise 15 members of Ascomycota; 4 members of Basidiomycota; and 1 member of Microsporidia. While Protist sequences are derived from 4 members of Apicomplexa; 4 members of Euglenozoa; 2 members of Entamoebidae; and one of each of Diplomonadida and Parabasalidea.

3.4.2 Performance of the sequence analysis workflows using Microbase

Microbase has facilitated multiple bioinformatics tools to be distributed across 74 desktop cluster machines at Newcastle University. Eight responders were developed during the study in order to analyse the large number of proteins sequence in an automatic manner. The overall time spent to

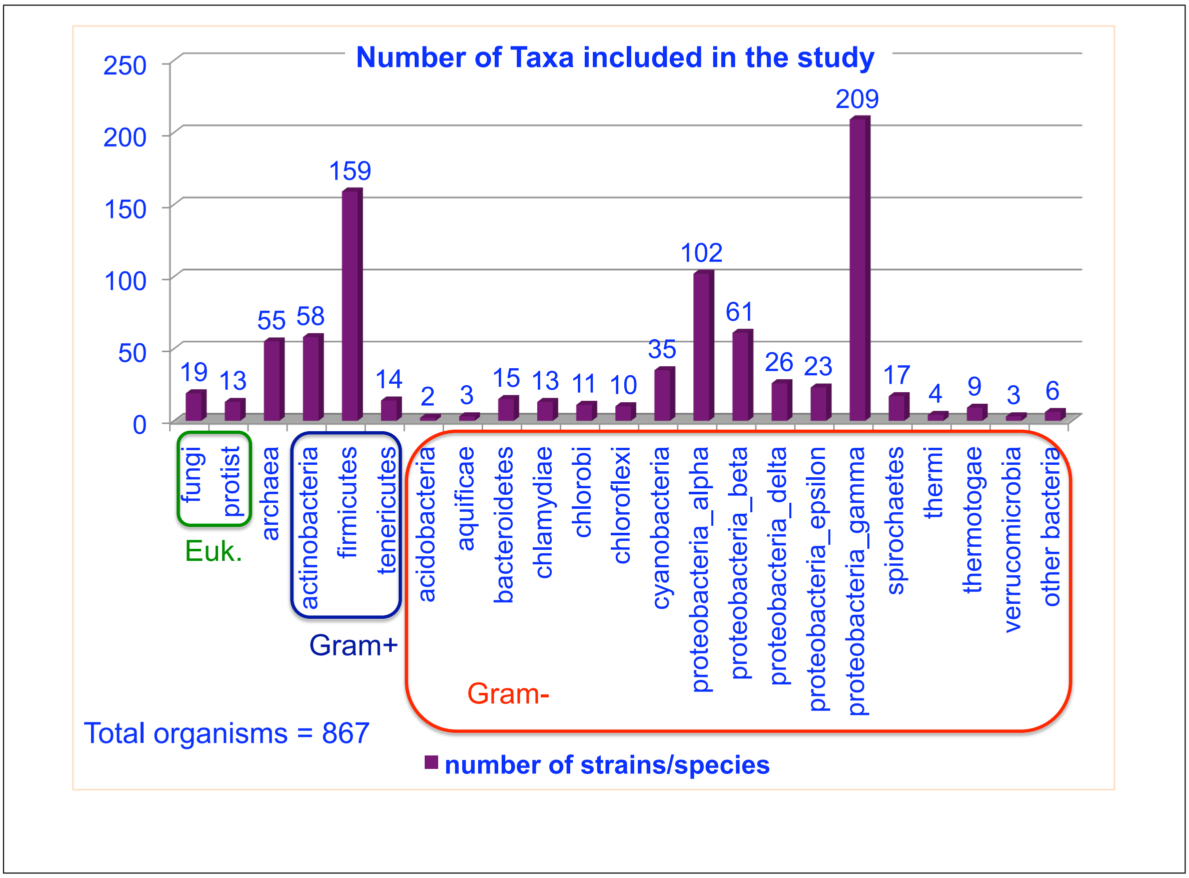
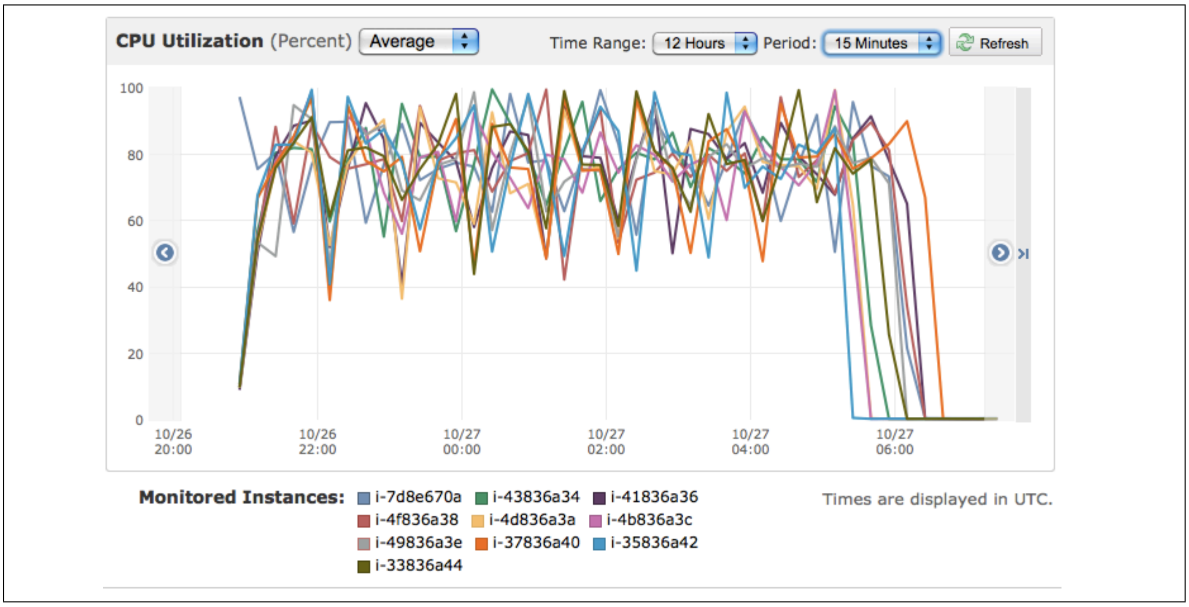


Figure 3.12: A summary of the number of taxa whose proteomes were included in the GenomePool database, with respect to their taxonomic groups. In total, the study contains genomes and corresponding proteomes from 867 microorganisms ranging from prokaryotes to eukaryotes. The number above each bar represents the total number of microorganisms species or strains whose genome sequence data were retrieved from RefSeq and stored in the GenomePool database. Taxa were grouped by high-level taxonomic classifications. Organism classes were also grouped with respect to their cell surface structures and their evolutionary distance in the global phylogenetic tree shown in [Ciccarelli *et al.*, 2006] (Eukaryotes, Archaea, Gram-positive and Gram-negative bacteria). Other bacteria include Dictyglomi, Elusimicrobia, Fusobacteria, Nitrospirae and Planctomycetes.

analyse protein sequences was reduced by using Microbase. Microbase therefore provides a significant advantage over the use of a single desktop machine to perform the various analysis processes. However, the extent of the speedup achieved varies between applications and primarily depends on the computational usage pattern of the application. For example, the Microbase FileScanner responder, which scans an FTP site and transfers new files to the resource system does not perform any CPU intensive work, and therefore does not speed up dramatically when parallelised. Similarly, the GenomeParser responder which reads GenBank file and parses sequence information into the database does not require high level of computational usage. The computational demands of these two responders is relatively low compared to other tools which require high computational demands, such as InterProScan or BLASTP (see Figure 3.13). A summary of the performance of all the responders is shown in Table 3.3. 3,153 GenBank-formatted genome files (jobs processed by the GenomeParser responder) describing the genomes of 867 microorganisms were parsed into the GPDB, resulting in 3,021,490 protein sequences. These protein sequences were the input data for the analysis pipelines (see Table 3.2). Using Microbase with 40 worker nodes, the process of populating the GPDB with complete genome sequences from 867 microorganisms from the three domains of life took 26 hours. After the GPDB had been populated, computational works for the TMHMM, SignalP, LipoP, InterProScan, BLASTP-pairwise and BLASTP-RefSeq tools could be processed in parallel. The responders responsible for executing these tools split the 3 million protein sequences into more manageable blocks of 100-1,000 sequences. The exact size of the blocks are responder-specific, depending on a particular tool limitations or compute time required to process a block of sequences. For example, the SignalP tool has a limit of 2,000 sequences and 200,000 amino acids allowed per execution[12], which if exceeded, results in a crash of the execution. The SignalP responder was set to execute 150 sequences per job to ensure a successful execution. The InterProScan responder was configured to process 100 sequences per job. InterProScan does not have a limitation on the number of sequences per input file, but the 100 sequence limit was applied in order to allow InterProScan jobs to complete within a reasonable amount of time (approximately 1 hour) without anticipated interference from regular users of the cluster machines. In total 101,943 compute jobs were produced by six sequence analysis responders. The hundred-thousand analysis jobs running different bioinformatics tools were assigned to 27-74 worker nodes on the Condor Grid system [Frey *et al.*, 2002] at Newcastle University, as well as the Amazon Cloud computing resource [Xu, 2010]. By exploiting high-throughput distributed computing resources, all the jobs were successfully completed within 2 months and all analysis results produced were stored in structured databases associated with each re-

[12] `http://www.cbs.dtu.dk/services/SignalP/`, accessed 20th October 2010

sponder. This duration includes the CPU usage for tool execution as well as additional time required for Microbase custom tasks such as input and output file management and automated software installation. The total ‘wall clock’ time spent for each responder to complete all jobs and an estimated total CPU usage time (computing time) for an ordinary machine to complete all the jobs is summarised in Table 3.3. Overall, InterProScan processes makes the most advantage of the distributed computing system: 5 years of computational time required to process 3 million sequences was reduced to 16 days of ‘wall clock’ time. Likewise, for BLASTP work, the amount of active time for processing approximately 26,000 BLASTP-pairwise and BLASTP-refseq jobs was also reduced significantly. The less CPU intensive programs such as LipoP, TMHMM and SignalP required 4 hours, 11 hours and 31 hours, respectively to complete the analysis on 27-74 cluster machines.



CPU Utilization (Percent) Average

Time Range: 12 Hours

Period: 15 Minutes

Figure 3.13: Average CPU usage of 10 Amazon EC2 machines running BLASTP jobs over a 12-hour period. Lines denote CPU usage (%) of worker nodes during that period.

3.4.3 Protein domain organisation prediction results

Protein sequences that contain known features or signatures predicted by InterProScan were identified. The results were obtained from the protein domain recognition pipeline, in particular, from the InterProScan result database produced by the InterProScanProcessor responder (see Section 3.3.1). The proportions of proteins carrying at least one known signatures to all the proteins included in the study was computed for every organism. The average of these proportions were calculated and summarised into taxonomic groups (see Figure 3.14). Protein carrying domain(s) were counted based on the entries of protein domain from InterPro database excluding the highly repetitive regions predicted by Seg [**?**]. For all taxonomic groups, an average of 76% of the sequences were predicted to have at

Table 3.3: Timing information of the Microbase responders developed for this project. 'Total CPU usage time' or total computing time shows the estimated time that a typical desktop computer might spend processing a particular task. A typical machine in this case refers to a desktop computer with specifications similar to the following: Intel core 2 (6300) duo 1.86 GHz CPU, 2GB memory. The 'total active time' column refers to the amount of time spent on software installation, input/output file management as well as job computation time and represents the 'wall clock' time take to complete each task as would be measured by a user with a stopwatch. For the FileScanner, GenomeParser and TMHMM responders, one Microbase job is created for each genome file. For the other responders, the number of jobs is determined by the responder-specific setting for the number of protein sequences allowed per job. '-' indicates that all the protein sequences annotated in a particular genome fragment file formed a single job.

Responder	**Total jobs**	**Maximum protein sequences permitted per job**	**Average time for a successful job execution (mins)**	**Total CPU usage time**	**Total active time**	**Average number of machines used**
FileScanner	3,153	-	0.01	27.99 mins	26 hrs	19
GenomeParser	3,153	-	0.32	16.67 hrs	26 hrs	40
TMHMM	2,892	-	1.88	3.78 days	11 hrs	74
SignalP	41,091	150	0.08	2.27 days	1 day 7 hrs	37
LipoP	2,941	1,500	0.04	1.93 hrs	4 hrs	27
InterProScan	31,924	100	68.12	1,510.15 days	16 days 15 hrs	60
SCL-BlastP	2,900	-	0.34	16.67 hrs	3 hrs	45
Hmmer-m60-like	2,900	-	0.06	2.78 hrs	2.5 hrs	48
BlastP-pairwise	22,801	-	1.44	22.81 days	17 days 5 hrs	40
BlastP-refseq	2,942	200	81.38	166.26 days	7 days 11 hrs	40

least one known sequence signature. Protein sequences from protists have the lowest average number of the fraction of proteins with known signatures (around 58%). The two organisms in the protist dataset with the lowest fraction of proteins of known signatures are *T. vaginalis* G3 and *Giardia lamblia* ATCC 50803 with proportions of 30% and 43%, respectively. The notably low fractions of proteins carrying known sequence signatures in the *T. vaginalis* proteome data set might be because this organisms has a relatively large proteome size (59,518 gene-coded protein sequences) as a result of a recent massive expansion of gene families [Carlton *et al.*, 2007]. However, it also reflects how little is known about the proteomes of these two vertebrate mucosa pathogens.

3.4.4 Extracytoplasmic protein identification results

The prediction results of extracytoplasmic proteins from proteins derived more than 800 microbial proteomes were summarised in this section. The results were derived from the extracytoplasmic protein prediction pipeline described previously (see Section 3.3.2). For each responder in the pipeline, the results generated by the associated bioinformatics tools were stored in a responder-specific structured database. The proportions of proteins predicted to have an alpha helical transmembrane, Sec signal peptides and lipoprotein signal peptides were estimated (see Figures 3.15, 3.16, 3.17.).

TMHMM prediction results

The prediction results produced by the TMHMMProcessor responder were stored in the TMHMM results database. In total, 698,134 sequences were predicted positive by TMHMM. Proteins carrying an alpha-helix transmembrane region are found in a range of 13-34% for bacterial proteomes, 16-26% for archaeal proteomes, and 14-22% for protist and fungi proteomes with some outliers. *Trichomonas vaginalis G3* appears to have the lowest proportion of proteins carrying alpha-helix transmembrane regions due to the massive proteome size mentioned earlier (see Section 3.4.3).

SignalP prediction results

563,941 sequences had positive SignalP predictions. Members of Proteobacteria show the most heterogeneity in the fractions of proteomes carrying Sec signal peptides, with an average of approximately 22%, a minimum of 2% and a maximum of 37%. As shown in Figure 3.16, the phylum Proteobacteria contains several outliers compared to other groups. The presence of these outliers could be a result of the total number of Proteobacteria in the analysis being markedly higher (n=420)

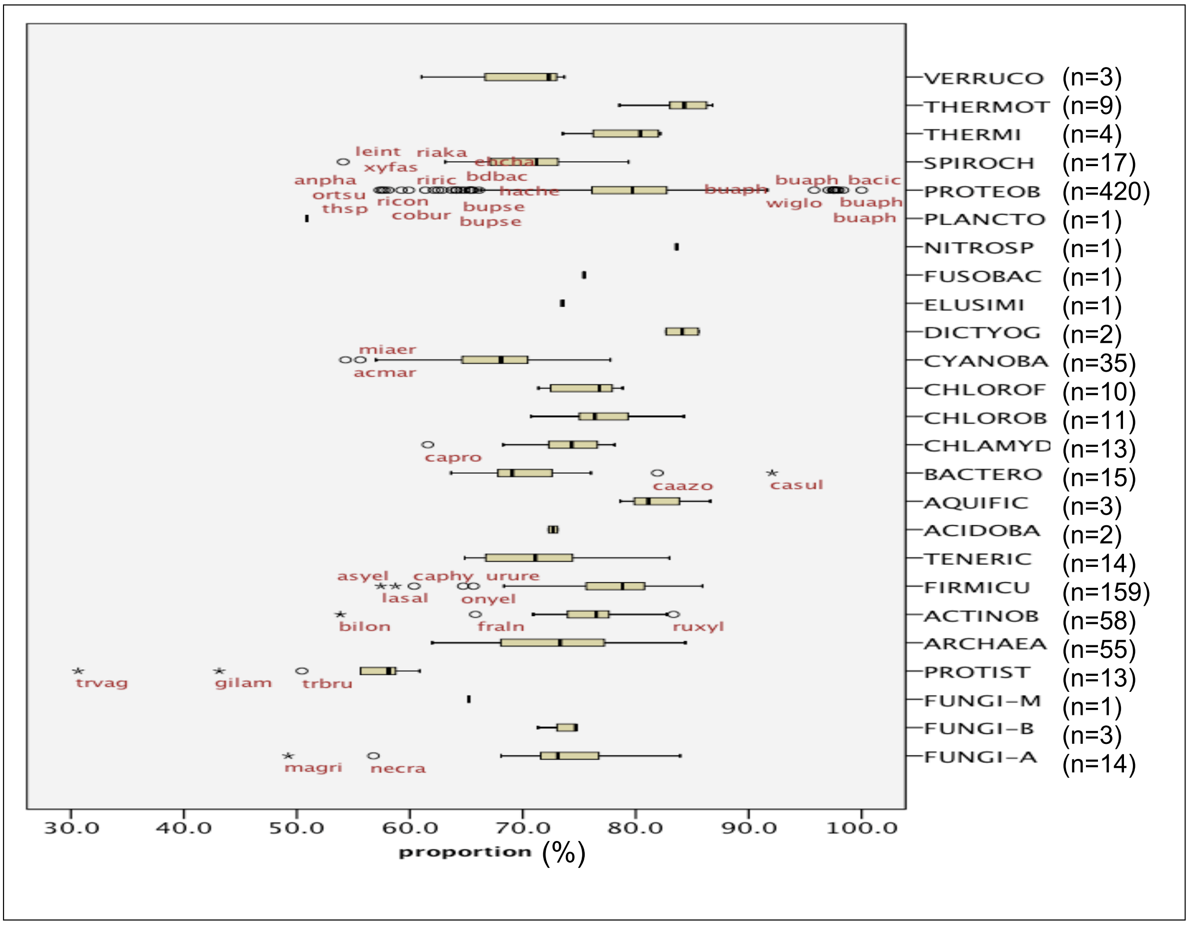


Figure 3.14: A boxplot displaying proportion of proteomes carrying known protein signatures from InterPro database. The vertical axis contains the boxplots for different taxonomic groups. The horizontal axis represents the average proportion of known protein signatures present in the proteomes of each taxonomic group. InterPro protein signatures were predicted by InterProScan. The proportion of proteins carrying at least one known signatures to all the proteins included in the study was computed for every organism. The resulting proportions were plotted with respect to organism taxonomic groups. The lower and upper edges of each box indicate the 25th and 75th percentiles, respectively, of the values found in a particular taxonomic group. The vertical line in each box indicates the median value of the data. The ends of the horizontal lines (whiskers) indicate the minimum and maximum data values. The whisker extends to a maximum of one quarter of the data unless outliers are present. Outliers are shown by open circles. Asterisks denote extreme outliers. n= number of taxa, leint= Leptospira interrogans serovar Lai str. 56601, riaka= Rickettsia akari str. Hartford , ehcha= *Ehrlichia chaffeensis* str. Arkansas, anpha= *Anaplasma phagocytophilum* HZ , riric= *Rickettsia rickettsii* str. Iowa, bdbac= *Bartonella bacilliformis* KC583 , ortsu= *Orientia tsutsugamushi* str. Ikeda , ricon= *Rickettsia conorii* str. Malish 7 , bupse= *Burkholderia pseudomallei* , hache= *Hahella chejuensis* KCTC 2396 , thsp= *Thauera sp.* MZ1T , cobur= *Coxiella burnetii* RSA 331 , buaph= *Buchnera aphidicola*, bacic= *Baumannia cicadellinicola* str. Hc (Homalodisca coagulata) , wigio= *Wigglesworthia glossinidia* endosymbiont of *Glossina brevipalpis*, miaer= *Microcystis aeruginosa* NIES-843 , acmar= *Acaryochloris marina* MBIC11017 , capro= *Candidatus Protochlamydia* amoebophila UWE25 , caazo= *Candidatus Azobacteroides* pseudotrichonymphae genomovar. CFP2 , casul= *Candidatus Sulcia* muelleri GWSS , asyei= *Aster yellows* witches'-broom phytoplasma AYWB , caphy= *Candidatus Phytoplasma* australiense , urure,lasal=*Lactobacillus salivarius* UCC118 plasmids, onyei= Onion yellows phytoplasma OY-M , bilon=*Bifidobacterium longum* DJO10A plasmid pDOJH10L, fraln= *Frankia alni* ACN14a , ruxyl= *Rubrobacter xylanophilus* DSM 9941 , trvag= *Trichomonas vaginalis* G3 , gilam= *Giardia lamblia* ATCC 50803 (*Giardia intestinalis* ATCC 50803), trbru= *Trypanosoma brucei* TREU927 , magri= *Magnaporthe grisea* 70-15 , necra= *Neurospora crassa* OR74A.

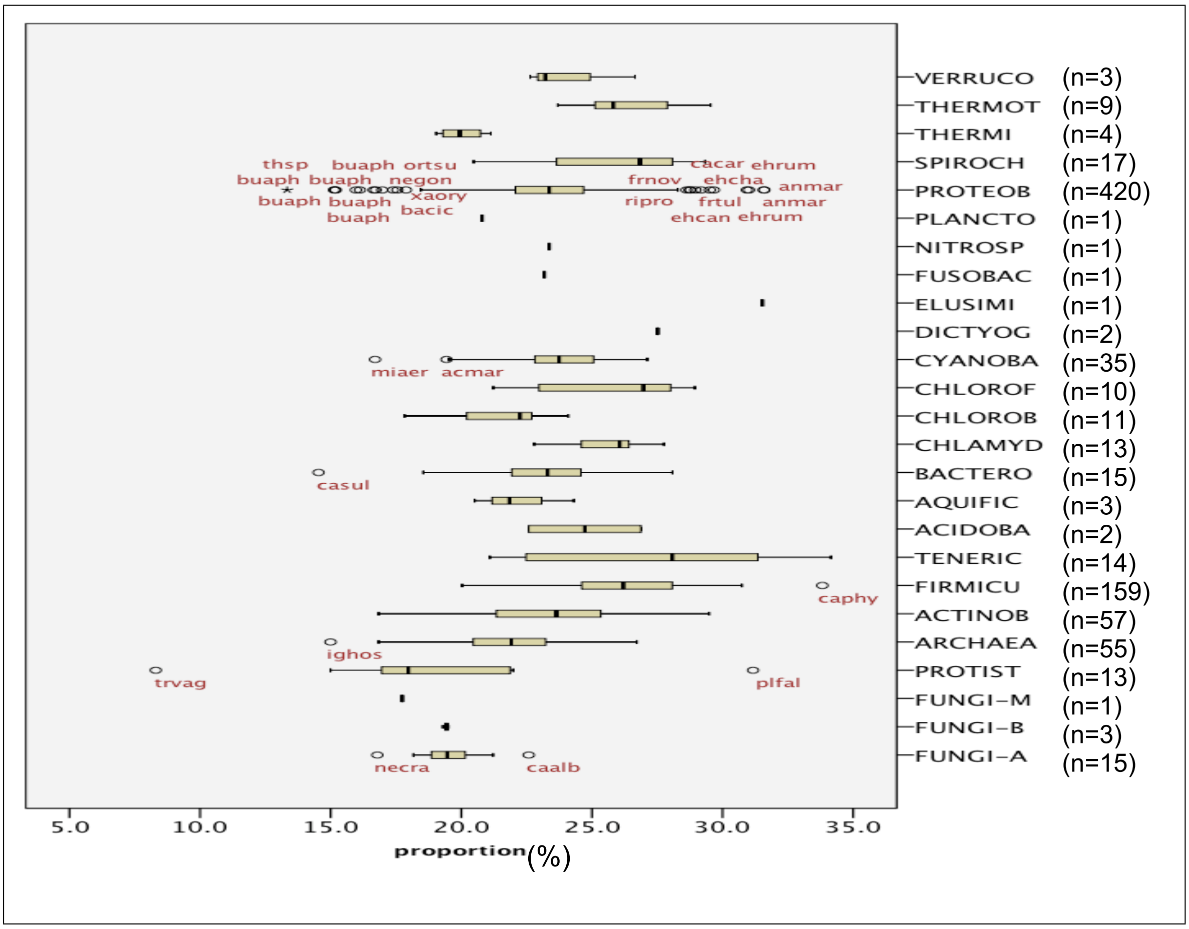


Figure 3.15: A boxplot displaying proportion of proteomes carrying alpha helical transmembrane domains in different phyla. The transmembrane proteins were predicted by the TMHMM tool. The interpretation of a boxplot is described in Figure 3.14. n= number of taxa, Thsp= *Thauera sp.* MZ1T, buaph=*Buchnera aphidicola* str. 5A (*Acyrthosiphon pisum*), ehcha= *Ehrlichia chaffeensis* str. Arkansas, frnov=*Francisella novicida* U112 , frtul= *Francisella tularensis* subsp. Holarctica, anmar= *Anaplasma marginale* str. Florida, ehrum= *Ehrlichia ruminantium* str. Welgevonden, str. Gardei, ripro= *Rickettsia prowazekii* str. Madrid E, miaer= *Microcystis aeruginosa* NIES-843, acmar= *Acaryochloris marina* MBIC11017, casul= *Candidatus Sulcia muelleri* GWSS, clpha= Clostridium phage phiSM101, caphy= *Candidatus Phytoplasma mali*, ighos= *Ignicoccus hospitalis* KIN4/I, trvag= *Trichomonas vaginalis* G3, plfal= *Plasmodium falciparum* 3D7.

than other phyla. Members of Proteobacteria also cover a broader range of microbial life styles, for example, free-living, marine, terrestrial, host-dependent, or intracellular pathogens.

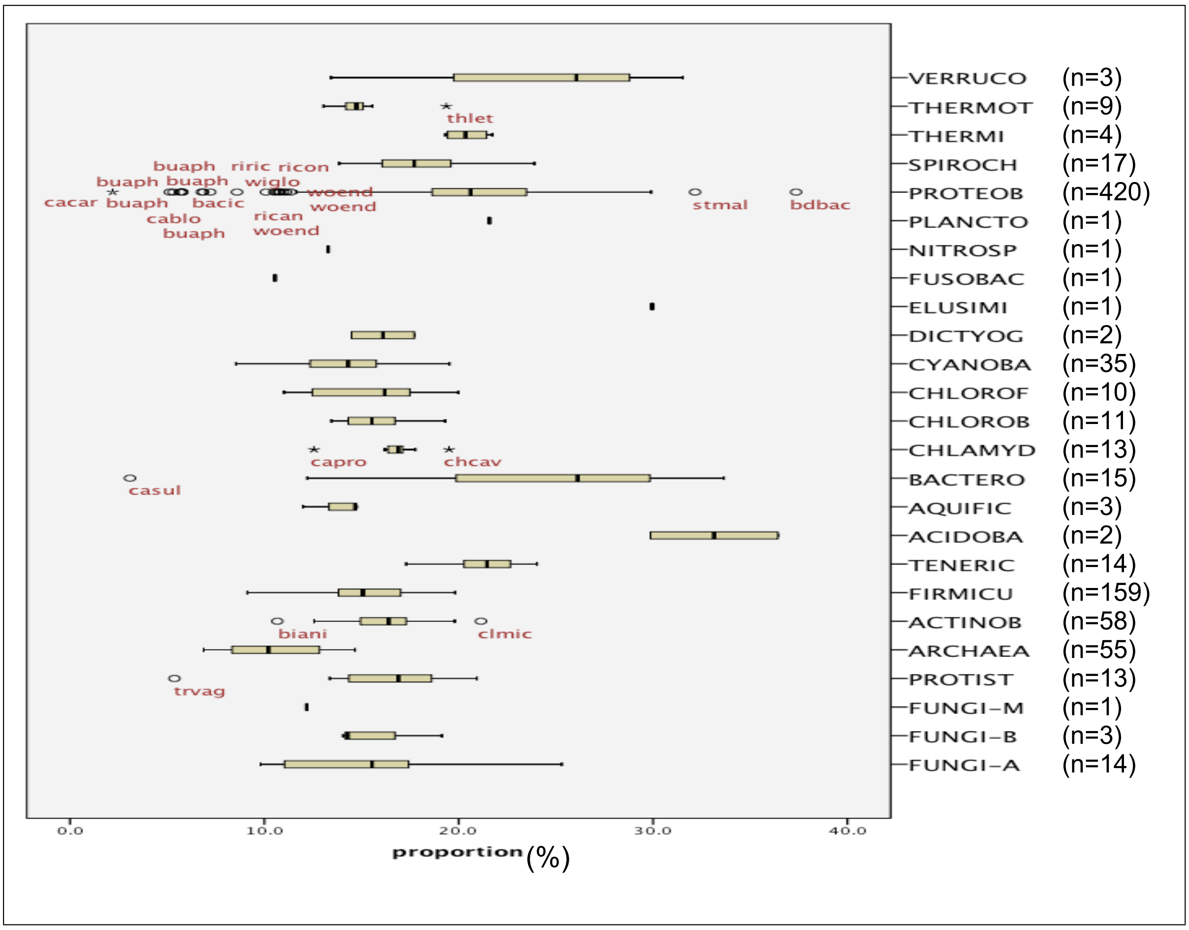


Figure 3.16: A boxplot displaying proportion of proteomes carrying the classical Sec signal peptides in different phyla. Sec signal peptides were predicted by SignalP. The interpretation of a boxplot is described in Figure 3.14. n= number of taxa, thlet= *Thermotoga lettingae* TMO, buaph= *Buchnera aphidicola*, bacic= *Baumannia cicadellinicola* str. Hc (Homalodisca coagulata) , cablo= *Candidatus Blochmannia* floridanus , riric= *Rickettsia rickettsii* str. 'Sheila Smith', ricon= *Rickettsia conorii* str. Malish 7, rican= *Rickettsia canadensis* str. McKiel, woend= *Wolbachia endosymbiont*, wigio= *Wigglesworthia glossinidia* endosymbiont of *Glossina brevipalpis*, stmal= *Stenotrophomonas maltophilia* R551-3, bdbac= *Bdellovibrio bacteriovorus* HD100, capro= *Candidatus Protochlamydia* amoebophila UWE25, chcav= *Chlamydophila caviae* GPIC, casul= *Candidatus Sulcia* muelleri GWSS, biani= *Bifidobacterium animalis* subsp. lactis AD01, clmic= *Clavibacter michiganensis* subsp. michiganensis NCPPB 382, trvag= *Trichomonas vaginalis* G3.

LipoP prediction results

63,468 sequences were positively predicted as lipoproteins by LipoP. The proportion of putative lipoproteins found in archaea ranges from approximately 0.03-2.9%, whereas bacterial proteomes appear to have much wider range of 0.3 to 12.6%. The phylum Bacteriodetes has the widest range of the proportion of lipoprotiens (1.3-12.5%) followed by Spirochaetes (1.7-7.8%) with an outlier. The results support the finding of Bendtsen 2005 [Bendtsen, 2005] showing that members of the phylum

Bacteriodetes exports more lipoproteins than other phyla.

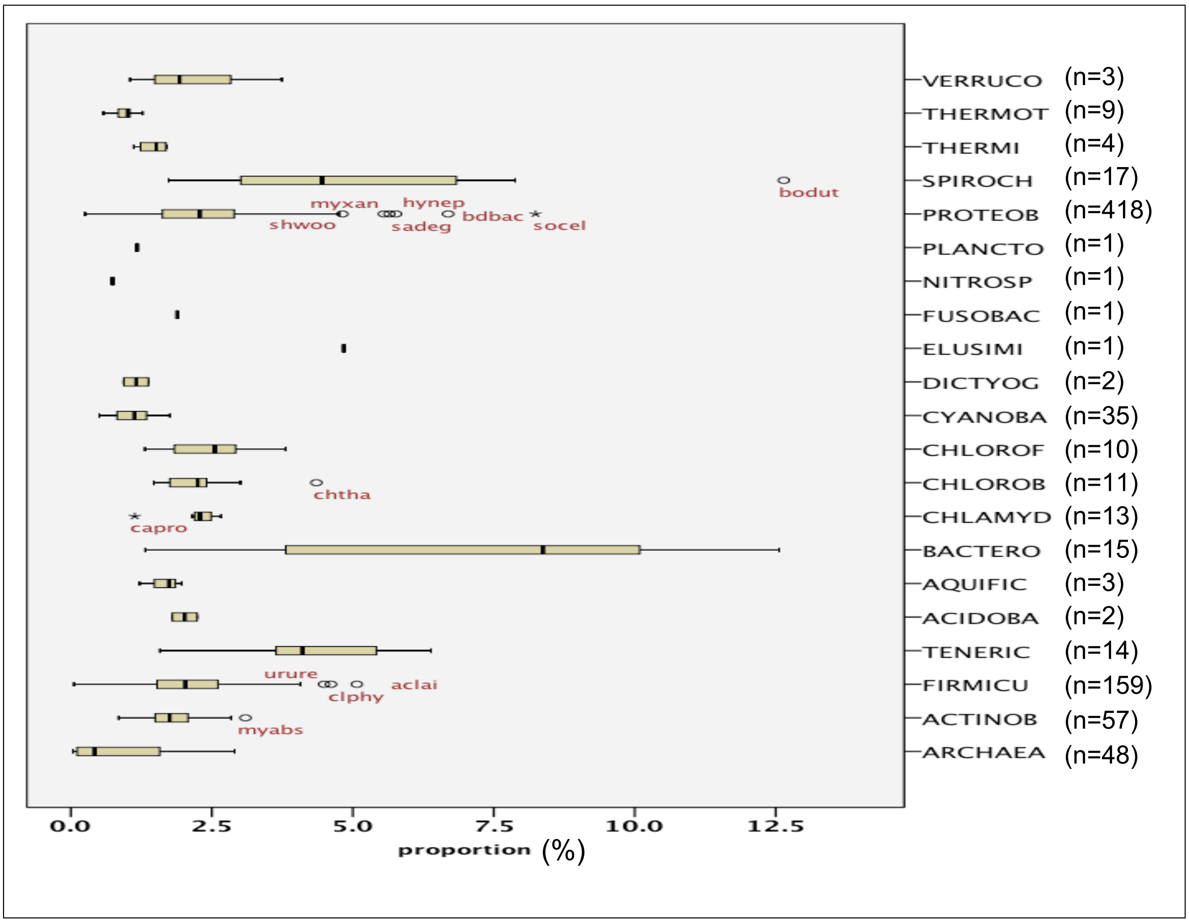


Figure 3.17: A boxplot displaying proportion of proteomes carrying lipoprotein signal peptides in different phyla. Lipoproteins were predicted by LipoP. The interpretation of a boxplot is described in Figure 3.14. n= number of taxa, myabs=*Mycobacterium abscessus*, capro=*Candidatus Protochlamydia amoebophila* UWE25, chtha=*Chloroherpeton thalassium* ATCC 35110, ighos=*Ignicoccus hospitalis* KIN4/I, urure=*Ureaplasma urealyticum* serovar 10 str. ATCC 33699, clphy=*Clostridium phytofermentans* ISDg , aclai=*Acholeplasma laidlawii* PG-8A, socel=*Sorangium cellulosum* 'So ce 56', bdbac=*Bdellovibrio bacteriovorus* HD100, hynep=*Hyphomonas neptunium* ATCC 15444, sadeg=*Saccharophagus degradans* 2-40, myxan=*Myxococcus xanthus* DK 1622, shwoo=*Shewanella woodyi* ATCC 51908, bodut=*Borrelia duttonii* Ly.

3.5 Discussion

To date, more than 1,000 completed genome entries have been made available in the GenBank database and new completed sequences or updated versions of existing sequences are being added on a daily basis. The bioinformatics workflows implemented using Microbase are capable of detecting when new sequences are released, downloading and parsing data into the structured database. The system also enables the processing of these sequences through a variety of tools.

The benefit of using a system such as Microbase is that each step is run automatically, without the need for human intervention. The system manages various aspects of running computational analyses such as automatically distribute of computational tasks to available worker machines, installing necessary software, staging input data files, executing the necessary bioinformatics software and finally managing the output data. When time spent for management and program execution steps are taken into consideration, such analytical processes could take several years if performed manually on a typical desktop computer. By using Microbase, a computationally intensive task can be split into several easier to manage chunks and can be distributed across several machines to execute in parallel. This feature significantly reduces the amount of chronological time taken to run compute-intensive analysis tasks on large amounts of data such as InterProScan and BLASTP pairwise all-vs-all homology searches on all 3,021,490 protein sequences in the GPDB. Other advantages include the fact that new sequences can be added into the system at any time and triggering further secondary or tertiary analyses. New results can be generated through the addition of new genomes and incorporated incrementally without recomputing the whole data set.

Although developing the responder-based system requires additional programming effort beyond simply automating a set of commands via scripts, the approach provides a highly modular system that allows the independent responders to be re-used. Several of the responders developed during this project are currently in active use by other bioinformatics research projects, such as AptaMEMS-ID [McNeil *et al.*, 2010]. The database components of the system are also reusable. In addition, the system also facilitates storage of the results (plain-text) in a structures form in relational databases enabling efficient querying. The output data from each tool can be integrated using a standard approach. In this project, the analysis results can be published as a pre-computed protein subcellular localisation prediction dataset, allowing large scale comparative genomics of extracytoplasmic proteins across the three domains of microbial organisms.

High-throughput bioinformatics workflow on Cloud VS Grid computing

High-throughput bioinformatics analysis workflows often require a reliable database server for storing primary sequence data, analysis results, as well as metadata generated by the system itself. In contrast, reliable worker machines are not required, but provide benefits in terms of increased efficiency. The greater the number of reliable machines available to process computing tasks, the more benefit can be gained from the Grid-based high-throughput computing system. Microbase allows the use of both Cloud and Grid computing resources.

The Condor Grid computing resource at Newcastle University are desktop machines which are available when no user is logged on or the CPU load is low. Microbase jobs running on these machines are terminated as soon as the machine is required by other users. Therefore, the jobs best suited to the Condor resource are short-running jobs since long-running jobs might be terminated before they complete. The termination of a job before completion effects the progress of that job. In contrast, the Cloud resource consists of dedicated hardware, and is guaranteed to be available at all time. Cloud-based machines are therefore suitable for long-running jobs without interruption. Moreover, Cloud hardware usually has higher specification in term of disk space, memory and network bandwidth than typical desktop machines. Therefore, Cloud-based machines are ideally suited to jobs that require a considerable amount of data staging, for example large BLAST database files.

The use of a high-throughput computing framework was shown to reduce the time required for the analysis of a large-scale sequence data set significantly. Running InterProScan on 100 protein sequences using an average desktop computer can take up to an hour (average 45 minutes) depending on the length of the input sequences. Using Microbase, one InterPro job responsible for an analysis of 100 protein sequences took 1 hour and 8 minutes on average excluding the handling time spent before and after a job execution (Table 3.3). Since worker nodes in the Condor Grid computing environment are shared with other users, long-running jobs may be frequently interrupted, and therefore take much longer to complete than an equivalent job executing on a dedicated machine. In Microbase, additional time is spent preparing a worker node. Such preparation processes include software installation, input file retrieval and output file transfers. Including the preparation and file transfer processes, the total node active time for processing all the 31,924 InterProScan jobs on an average of 27 machines in the Newcastle University's cluster machines was 16 days and 15 hours. It would therefore take approximately 5 years to analyse the 3 million proteins in this project using InterProScan on a single desktop machine that is not shared by several users. Microbase has been shown to speed up a large-scale protein sequence analysis as all the project-related responders' jobs concerning GPDB construction and sequence analyses were completed within 2 months.

Challenges in integrating bioinformatics tools into a high-throughput computational workflow.

Numerous protein subcellular localisation prediction tools are available and new approaches are being developed to increase the accuracy of these tools. Different approaches and methodologies have been established with the main aim to increase the accuracy level of the predictions. In order for a tool to be integrated into Microbase, a machine-usable interface, such as a command line (downloadable as a standalone version) or web service interface is required. The vast majority of analysis tools needed for this project were available in one or both of these forms. However, some tools such as BaCelLo [Pierleoni *et al.*, 2006] are only available online via a web-based human-accessible interface. While it may be possible to implement an adaptor to utilise the tool, such methods are often difficult to implement and maintain, and are often unreliable. Therefore, although BaCelLo is publicly available, and has been shown to slightly outperforms other tools such as TargetP, it is currently not an option to incorporate it into a Microbase pipeline.

PSORTb version 2.0 is an example of a bioinformatics tool that, in theory provides a desirable set of functionality to predict protein sublocalisation, but for practical reasons was not feasible to use in Microbase. The tool required several dependency libraries, which in turn required administrator access to configure and install. Many of the machines available for processing the analysis at Newcastle University are production machines running a standard set of software. Therefore, it was not feasible to make major alterations to these machines in case the PSORTb requirements interfered with the standard campus software.

3.6 Conclusions

Several protein sequence analysis workflows that employ an event-based distributed computing system (Microbase) has been successfully developed. The resulting system allows large-scale computationally intensive bioinformatics tasks to be performed simultaneously on a number of machines. The workflows constructed are capable of analysing large amount of genomics data and executing a variety of bioinformatics tools. The computational tasks generated by the workflows were simultaneously distributed to CPUs on the Amazon Cloud computing system and Newcastle University cluster machines. The speed of running this large-scale sequence analysis was improved by about 60 times over the time that would have been needed to execute the analysis on a single computer. All the analysis finished within a few months by using a set of 40-80 machines rather than the estimated time of 5 years by using a single desktop computer. By exploiting high-throughput computing, where multiple machines work in parallel, the computational part of this project was completed in 2 months.

The Microbase system is highly modular, allowing workflows to be extended with new tools in a relatively straightforward manner. The workflow in this study can be reused in a larger set of genomes or extended to execute other standalone bioinformatics tools with minimal effort. Not only can the analysis be re-executed if necessary at any time, but also new data can be automatically integrated with an existing data set without re-computing or repeating already completed analyses. The working Microbase system automates the process of producing bioinformatics datasets, allowing a biologist to analyse the data without having to worry about developing computing infrastructure such as the installation and execution of software or output data integration. The bioinformatics workflows developed using a high-throughput computing framework have proven to be satisfactory for facilitating post-genomics studies requiring actions to be performed in a systematic and automated fashion. This approach should continue to be useful in a field where the amount of data to be analysed increases exponentially every few months.

Chapter 4

Computational Approaches to the Identification of Microbial Extracytoplasmic Proteins

4.1 Introduction

Microbial extracytoplasmic proteins including secreted, surface-associated, and transmembrane proteins (see Figure 2.8) mediate key processes underlying in host-microbe interactions of both mutualistic and parasitic microbial partners [Pallen and Wren, 2007] [Turnbaugh *et al.*, 2007]. The extracytoplasmic proteomes of pathogenic strains are known to play a role as virulence factors that mediate the pathogenesis in the host body [Dreisbach *et al.*, 2010]. These factors include proteins involved in, for instance, quorum sensing, cell adhesion, secretion systems and toxin production [Barczak and Hung, 2009] [Lebeis and Kalman, 2009]. Targeting microbial virulence therefore provides an alternative or complementary strategy to the development of antibiotic treatment [Clatworthy *et al.*, 2007] [Cegelski *et al.*, 2008]. Moreover, a pathogen's extracytoplasmic proteomes are potentially good targets for biomarkers for diagnostic assays and vaccine development [Pajón *et al.*, 2006] [Lee *et al.*, 2003] [Lin *et al.*, 2002]. A recent study of the surface proteomes of several different *Staphylococcus aureus* strains has shown a high degree of variability of the surface proteins between the strains [Dreisbach *et al.*, 2010].

One of the main focuses of this project was to perform an *in silico* functional analysis on microbial extracytoplasmic proteins in order to identify proteins that are important for the survival of microbes in the host body. The study includes a large number of proteins from multiple groups of microorganisms. In order to acquire a relatively high quality set of candidate extracytoplasmic proteins from approximately three million sequences derived from both prokaryotes and microbial eukaryotes, there was a need for an approach to computationally identify extracytoplasmic proteomes of microorganisms with various cell surface structures and chemical properties. Ideally, such an approach should provide a large number of potential extracytoplasmic protein candidates.

Numerous bioinformatics tools exist for predicting protein subcellular localisation, each of which has different benefits and flaws. The diversity of cell surface structures also complicates the prediction of protein location. It is therefore non-trivial to precisely identify extracytoplasmic proteins computationally since there are several mechanisms and pathways that target particular proteins through cell membranes (see Section 2.3.3). Different groups of organisms typically use different mechanisms of protein translocation through a cell membrane. These mechanisms are different depending on types of organisms and their membrane structures (see Section 2.3.1).

Many prediction tools have been developed for the identification of specific targeting signals, such as SignalP for prediction of N-terminal signal peptides [Dyrlovbendtsen, 2004] and TMHMM for the identification of alpha-helix transmembrane regions [Krogh *et al.*, 2001]. Several well-known protein-targeting signal predictors have been trained with a set of proteins from a limited group of taxa. For example, SignalP was trained with experimentally-verified cytoplasmic protein sequences and with sequences containing N-terminal signal peptides. Firmicutes and Gracilicutes protein sequences were used as Gram-positive and Gram-negative bacterial protein training data sets, respectively [Dyrlovbendtsen, 2004]. Some tools integrate several algorithms and techniques in order to achieve a high performance in the prediction of protein sublocalisation. Most of the tools have been designed to work with sequences from either eukaryotes or prokaryotes, but not both. PSORTb is a prediction tool that combines several algorithms in order to identify multiple targeting signal features and also to predict potential subcellular localisations for prokaryotic proteins. However, PSORTb was designed to emphasize positive predictive value over sensitivity (average at 96% positive predictive value, 64% sensitivity in the data set originally published in [Gardy, 2004] with the tool).

In this study, a workflow combining different programs, determination criteria, methods and strategies was developed. An initial aim of developing this identification workflow was to construct a system that can be used with primary amino acid sequence data from all three domains of life and provide a relatively high quality of prediction results. Another purpose of the workflow was to computationally generate an initial list of potential candidate extracytoplasmic proteins to serve the needs of a high-throughput sequence analysis in the post-genomics era. Further workflows were also constructed to distinguish surface-associated and secreted proteins from transmembrane and other proteins in periplasmic space.

This chapter includes a description of the approach used to generate an initial list of potential candidate extracytoplasmic proteins and the approaches used to identify cellular location of the extracytoplasmic proteins. The performance of the extracytoplasmic protein identification workflow was then evaluated. Finally, this chapter reports the application of the workflow to the *Bacillus* extracytoplasmic proteome. Extracytoplasmic protein domains identified as shared or being predominant among *Bacillus* species are also reported.

4.2 Materials and methods

Several existing high performance prediction tools were employed to identify putative extracytoplasmic proteins in a data set of three million protein sequences from both prokaryotes and microbial eukaryotes. A computational workflow incorporating these existing tools was constructed to automate the process of extracytoplasmic protein identification. This workflow was developed using the distributed-computing framework, Microbase (see Section 2.8.4).

The work in this chapter utilises data generated from the sequence analysis workflows described in Chapter 3 (Section 3.3.2, 3.3.3). This data mainly consists of prediction results from several well known bioinformatics tools, SignalP, LipoP, TMHMM, InterProScan and BLASTP, used to identify targeting-signals and other protein sequence signatures. The process of executing these tools on the input protein sequences and obtaining the analysis outcomes is also described in Chapter 3 (see Section 3.3). This section describes the strategy used to construct a workflow for the identification and classification of microbial extracytoplasmic proteins.

4.2.1 The extracytoplasmic protein identification workflow

The extracytoplasmic protein identification workflow was developed to generate a list of candidate proteins by considering the results from the analysis workflows described in Chapter 3 (Section 3.3.2, 3.3.3). In this workflow, the results from each sequence analysis tool employed were considered with respect to the types of organism that each tool was trained with, or suitable for. All positive predictions were stored in the in-house database developed as part of this project, PredExtDB (discussed in Section 3.2.4). The technical procedure of filtering results into the PredExtDB was described earlier in Chapter 3 (Section 3.3.3). Here, the inclusion criteria for the acquisition of putative extracytoplasmic proteins into the PredExtDB was described. Proteins were considered as putative extracytoplasmic proteins (i.e., stored in PredExtDB) if they had any of the following features:

- an alpha-helix transmembrane region(s) predicted by TMHMM;
- a N-terminal signal peptide cleavage site predicted by SignalP;
- a feature type signal peptidase II (SpII) predicted by LipoP;
- homology to any known experimentally-verified extracytoplasmic proteins from the data set derived from ePSORT using BLASTP searches with an e-value less than 1×10^{-9}. The length of the query protein must also be within a range of 80-120% of the experimentally-tested extracellular proteins;
- possession of any known functional domain typically exposed to extracellular space including the surface-associated domains listed in Table 3.1;
- possession of surface or membrane-associated GO terms listed in Table 3.1.

A summary of the strategy used to classify a protein as 'extracytoplasmic' is shown in Figure 4.1.

4.2.2 Extracytoplasmic protein classification workflows

The focus of this project originally was to study the microbial surface proteome and secretome (extracellular proteins; see Figure 2.8). Given the list of putative extracytoplasmic protein candidates from the previous workflow (see Section 4.2.1), extracellular proteins had to be differentiated from transmembrane proteins as well as other proteins in the peptidoglycan layer. Therefore, further computational workflows were constructed in order to classify the initial list of candidate extracellular proteins in more detail. Three workflows were designed to handle microbes with different types of cell surface structures: Gram-negative bacterial, Gram-positive bacterial and eukaryotic microbial cell surface. In this study, the workflow designed for Gram-positive bacterial proteins was also applied to proteins derived from Archaea. This strategy was used because archaeal cell surfaces typically comprise of one lipid bilayer membrane, similar to Gram-positive bacteria. A combination of results from several prediction tools was used to group proteins into classes depending on the consistency of the predictions. Putative extracellular proteins were classified into six different classes depending on targeting signals predicted by the tools described in the Section 3.3. These classes represent artificial categories of extracytoplasmic proteins taking into account the presence of evidence

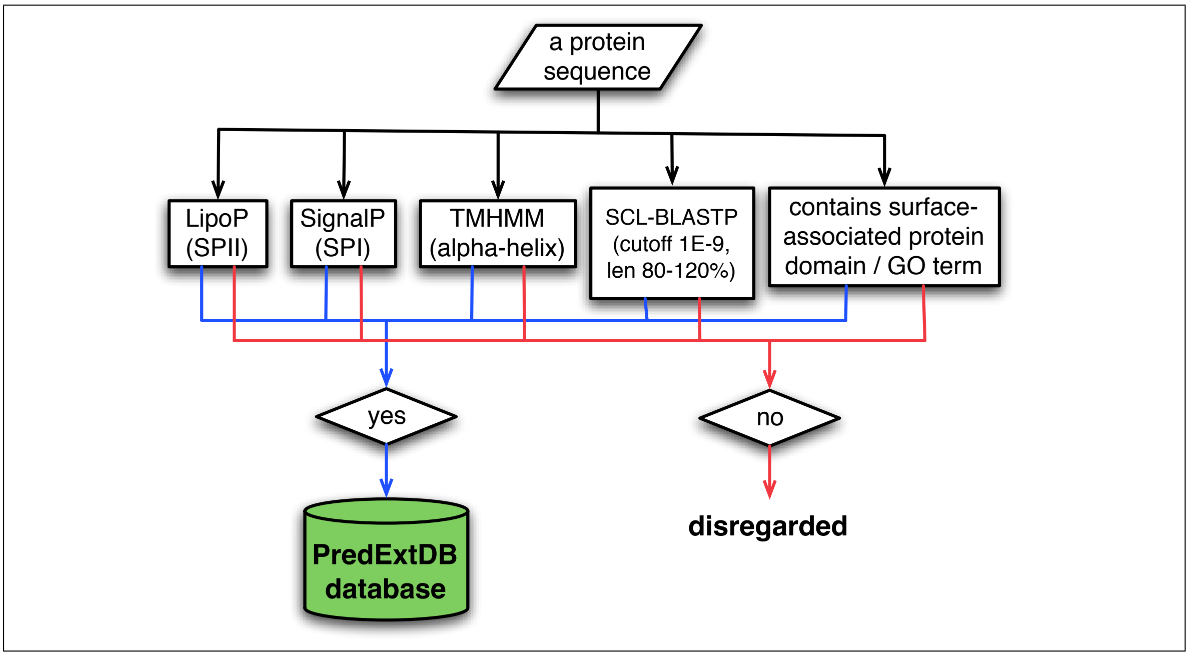


Figure 4.1: Flow chart for the identification of extracytoplasmic proteins. A protein sequence was considered as extracytoplasmic and added to a list of putative extracytoplasmic proteins if at least one of the criteria listed in the square box is evaluated to be true. The inclusion criteria are positive prediction results from LipoP, SignalP, TMHMM, SCL-BLASTP and surface-associated protein domains and Gene Ontology (GO) terms. Notes in brackets represent results considered as positive predictions. The BLASTP search was performed against the set of known experimentally-verified extracytoplasmic proteins obtained from ePSORTdb. Surface or membrane-associated protein domains and GO terms are listed in Table 3.1. All the results were stored in the predicted extracellular protein database (PredExtDB) developed as part of this project (see Section 3.2.4). SPII = Signal peptidase II, SPI = Signal peptidase I, SCL = subcellular localisation, len = length

supporting their localisation outside the cytoplasmic space. Therefore, these categories may not necessary reflect the actual cellular location of the proteins. The six classes of extracytoplasmic proteins are described in the following list. It is important to note that protein sequences were classified as a series of steps, and therefore each sequence can only belong to one class type. The classification process for a particular protein stops once a sequence is assigned to a class. The classes are listed beow.

- Multiple-TM protein: sequences with more than one alpha-helix transmembrane region predicted by TMHMM.
- One-TM protein: sequences with one predicted transmembrane segment located after the N-terminal targeting signal cleavage site predicted by SignalP or LipoP.
- Lipoprotein: sequences with SPII cleavage sites predicted by LipoP without a predicted transmembrane region after the cleavage site;
- Sec-pathway protein: sequences with SPI cleavage sites predicted by SignalP with no predicted transmembrane regions after the cleavage site. Proteins in this class can be regarded as being exported from cytoplasm, which can then become secreted or surface anchored proteins. GPI-anchoring proteins are potentially assigned to this class due to the presence of the signal peptide.
- Proteins with extracellular domains: sequences with predicted surface-associated protein signatures or GO terms predicted by InterProScan.
- Sequences homologous to verified extracytoplasmic proteins: sequences that are highly similar to experimentally verified bacterial extracytoplasmic proteins identified by BLASTP with an e-value cutoff of 1E-9 and whose sequence length is within 80-120% of the hit subject.

The first step of the classification workflow differentiates transmembrane proteins from secreted and surface proteins. The prediction results from the N-terminal targeting signal and transmembrane predictors from SignalP, LipoP and TMHMM were considered at this stage. TMHMM and SignalP have a well-known weakness resulting from their dependence on finding a region of hydrophobic residues to determine alpha-helix transmembrane regions and N-terminal signal peptides. This common recognition pattern between the tools leads to overlap between the two types of predictions [Lao *et al.*, 2002][Krogh *et al.*, 2001]. For example, the hydrophobic core of a signal peptide is frequently mistakenly predicted as a putative helix transmembrane segment by TMHMM. Likewise, SignalP reports many false positive results due to the exclusion of a transmembrane domain prediction [Menne *et al.*, 2000]. For this reason, results from the transmembrane topology and signal peptide prediction methods were combined to allow the differentiation of the true transmembrane protein from sequences with targeting signals and no predicted helix.

Combining the results from these algorithms requires several transformation steps. Firstly, sequences that are predicted to have more than one helix by TMHMM must be extracted. These filtered sequences were then marked as putative multiple-transmembrane (multiple-TM) proteins. Next, predicted N-terminal targeting signals and predicted one-helix sequences must be discriminated. Sequences with N-terminal targeting signals predicted by LipoP or SignalP were checked for the presence of any helix region located N-terminally to the signal peptide cleavage site. If no helix region exists, the sequence was classified as an N-terminal signal targeting protein: either a lipoprotein if identified by LipoP or a Sec-pathway protein predicted if detected by SignalP. If a helix segment was predicted after the signal peptide cleavage site, the sequence was classified as a putative one-transmembrane (one-TM) protein. The workflow developed in this project considered the prediction results from LipoP prior to taking into account the SignalP prediction results because LipoP was developed particularly to distinguish the SPII-cleaved proteins (lipoprotein) from the SPI-cleaved proteins [Juncker *et al.*, 2003].

The developers of the SignalP, LipoP and TMHMM tools focused on relatively limited sets of organisms considering the much wider range of taxonomic groups analysed in this project [Dyrlovbendtsen, 2004] [Juncker *et al.*, 2003] [Krogh *et al.*, 2001]. Relying on only the predictions from these algorithms for all the protein sequences included in this project therefore may not be optimal for the broad range of organisms used in this project. To maximise the number of true positive predictions of extracytoplasmic proteins, surface-associated protein domain annotations and GO terms identified by InterProScan were also incorporated. A list of the surface-associated protein domains and GO terms taken into account for this step is shown in Table 3.1. Sequences annotated with any of the surface-associated domains or GO terms were classified as proteins with extracellular domains.

The final stage of the workflow examined the sequences that were not yet filtered by any of the steps described earlier. This step considered the positive results from the SCL-BLASTP search (described in Section 3.3.2 and Section 3.3.3). The remaining sequences had no predicted targeting signals, transmembrane helices, or known surface-associated protein domains but were highly similar to known experimentally-verified extracytoplasmic proteins. These sequences have a BLAST e-value of $< 1E\text{-}9$ and the length of the sequences was within 80-120% of the experimentally-tested extracellular proteins.

The extracytoplasmic protein location classification results of proteins from the use of the workflow is discussed in detail in the following sections.

Approach for classifying Gram-negative bacterial extracytoplasmic proteins

The workflow described above was applied for the classification of Gram-negative extracytoplasmic proteins into six classes (Figure 4.2). This approach was applied to Gram-negative bacterial proteins or proteins from other non Gram-staining prokaryotes with an outer membrane. This group of organisms includes Bacteroidetes, Proteobacteria, Spirochaetes, Chlamydiae, Acidobacteria, Aquificae, Chlorobi, Chloroflexi, Cyanobacteria, Thermi, Thermotogae, Verrucomicrobia.

Approach for classifying Gram-positive bacterial and archaeal extracytoplasmic proteins

For Gram-positive bacterial proteins, the workflow classified the putative extracytoplasmic sequences into six classes (Figure 4.4). In this workflow, the results from SignalP were considered as an additional step in the lipoprotein identification in order to reduce false positives that may have been introduced by LipoP. In general, the length of a signal peptide of a lipoprotein is shorter than that of a Sec-type secretory protein [Juncker *et al.*, 2003] [Tjalsma *et al.*, 2000]. Since LipoP was trained with a set of known Gram-negative lipoproteins, but none from Gram-positive bacteria [Juncker *et al.*, 2003]. In our workflow for Gram-positive bacterial proteins, a protein was classified as a lipoprotein if there was a positive prediction by both LipoP and SignalP. The strategy here was to use the ability of SignalP to predict a hydrophobic region that is located prior to the cleavage site. In this workflow, LipoP was used to provide a prediction of a SPII cleavage site (c-region), while SignalP results were employed to ensure the presence of an N-terminal hydrophobic region (h-region). The cross-validation was performed by using SignalP prior classifying sequences as lipoproteins in order to increase the true positive prediction of Gram-positive lipoproteins (see Section 4.3.1).

During the course of this project, it was observed that LipoP sometimes misreports Gram-positive sequences as lipoproteins due to the presence of a potential c-region, but these sequences actually do not contain the h-region. Such sequences should not be classified as lipoproteins since the signal sequence can be characterised by the presence of the h-region as well as the c-region. For example, the *B. subtilis*' prephenate dehydratase[1] (EC:4.2.1.51) encoded by *PheA* gene, is involved in amino acid biosynthesis that would need to take place in the bacterial cytoplasm [Wipat *et al.*, 1996]. In addition, the protein was not recognised as an extracellular protein by the review on the secretome of

[1] `http://www.genome.jp/dbget-bin/www_bget?bsu:BSU27900`, accessed 20th May 2010

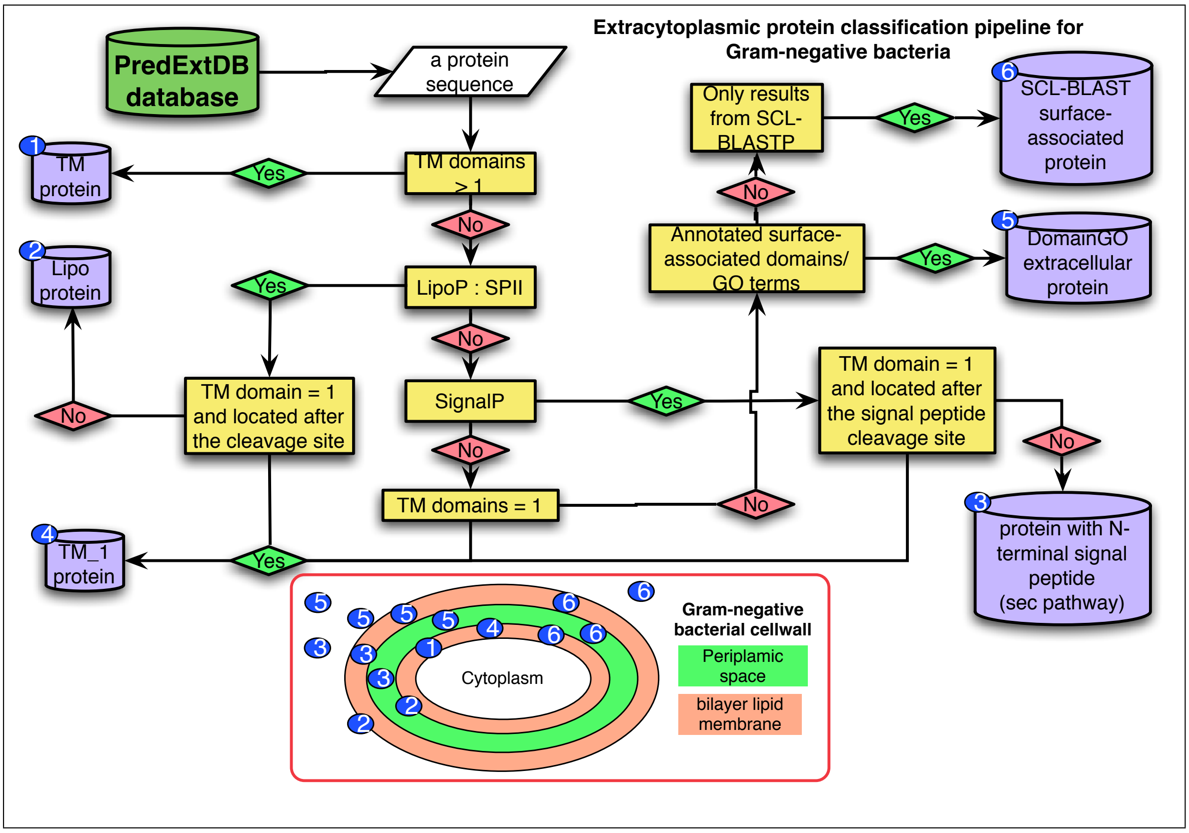

Extracytoplasmic protein classification pipeline for Gram-negative bacteria
PredExtDB database
a protein sequence
TM domains > 1
Yes
1 TM protein
No
LipoP : SPII
Yes
2 Lipo protein
TM domain = 1 and located after the cleavage site
No
SignalP
Yes
TM domain = 1 and located after the signal peptide cleavage site
No
3 protein with N-terminal signal peptide (sec pathway)
TM domains = 1
No
Yes
4 TM_1 protein
Annotated surface-associated domains/ GO terms
Yes
5 DomainGO extracellular protein
No
Only results from SCL-BLASTP
Yes
6 SCL-BLAST surface-associated protein
Cytoplasm
Gram-negative bacterial cellwall
Periplamic space
bilayer lipid membrane

Figure 4.2: ExCyt protein classification pipeline for Gram-negative bacteria. Putative ExCyt proteins were identified using the results from TMHMM, SignalP, LipoP, SCL-BLASTP (searching against a set of experimentally verified surface-associated proteins) and surface-associated protein domain/GO terms annotations. The set of predicted ExCyt proteins were retrieved from our analysis database. A protein was systematically classified into one of six different classes: 1) Transmembrane (TM) protein if more than two helices were identified; 2) Lipoprotein if the SPII cleavage site was predicted by LipoP without a TM located after the cleavage site; 3) Secreted protein via the Sec pathway if predicted positive by SignalP but either no TM domain was predicted, or a predicted TM domain was located N-terminally to the signal peptide (SP) cleavage site; 4) Protein with a single TM domain if a TM domain was identified without a SPI or SPII cleavage site or the TM domain was located C-terminally to the cleavage site; 5) Protein with surface-associated domains/GO terms if annotated as such; 6) SCL-BLAST Protein if a BLAST hit with an e-value < 1E-9 by the SCL-BLASTP analysis was present. Each putative ExCyt protein can only be classified into one of the six defined classes (the first classification that matches). Data storage for each class is shown in purple. Yellow squares represent processing steps. Yes/No decisions made for each step throughout the classification pipeline are highlighted in green and red, respectively. Arrows show direction of the workflow.

Bacillus subtilis carried out by Tjalsma *et al.* [Tjalsma *et al.*, 2004]. The *Bacillus* PheA protein has a positive LipoP prediction, however, no h-region was detected on the PheA protein using SignalP (see Figure 4.3). In this case, the result from SignalP showed no evidence for an h-region. Therefore, it can be concluded that PheA is less likely to be a lipoprotein candidate even though LipoP indicated the presence of SPII cleavage site on the sequence.

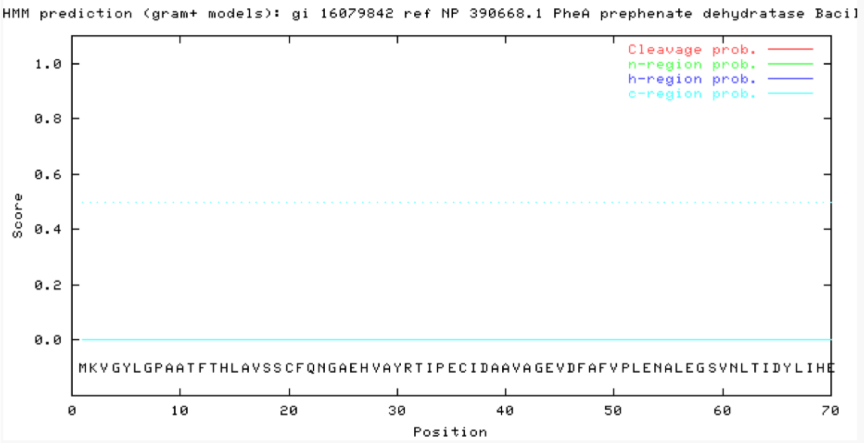


Figure 4.3: The graphical result from SignalP-HMM prediction on *B. subtilis*' prephenate dehydratase. This figure suggests an absence of hydrophobic region (h-region) on the N-terminal of protein sequence. However, the sequence is predicted positive by LipoP as a putative lipoprotein with a SPII cleavage site at amino acid position 19-20.

The Gram-positive lipoprotein predictions based on the combination of results from LipoP and SignalP are shown in the result section (Section 4.3.1).

To date, very few prediction tools are specifically designed to predict archaeal protein subcellular localisations and none of them work in a standalone or programmatically automate-able manner. When identifying of archaeal extracytoplasmic proteins, the workflow mainly relied on the same prediction tools and workflow developed for Gram-positive bacterial proteins. The same workflow was used for proteins from both prokaryotic groups due to the commonality between their cell surface. The overall structure of archaeal cell surfaces are similar to Gram-positive bacteria; they have a single plasma membrane with or without a cell wall, but lack an outer membrane and periplasmic space [Ellen *et al.*, 2010][Golyshina and Timmis, 2005]. A bioinformatics study of a subset of archaeal proteins with putative signal-peptides has suggested the characteristics of the signal peptides are more similar to bacterial signal peptides than eukaryotic ones [Bardy *et al.*, 2003]. Moreover, several studies have proposed the existence of lipoproteins in various archaeal species [Albers and Driessen, 2002][Kokoeva *et al.*, 2002][Mattar *et al.*, 1994]. Therefore, the workflow developed in this project for Gram-positive bacterial project was also applied to archaeal proteins as well as Gram-positive bacterial proteins. Gram-positive bacterial organisms here include Actinobacteria, Firmicutes and Tenericutes.

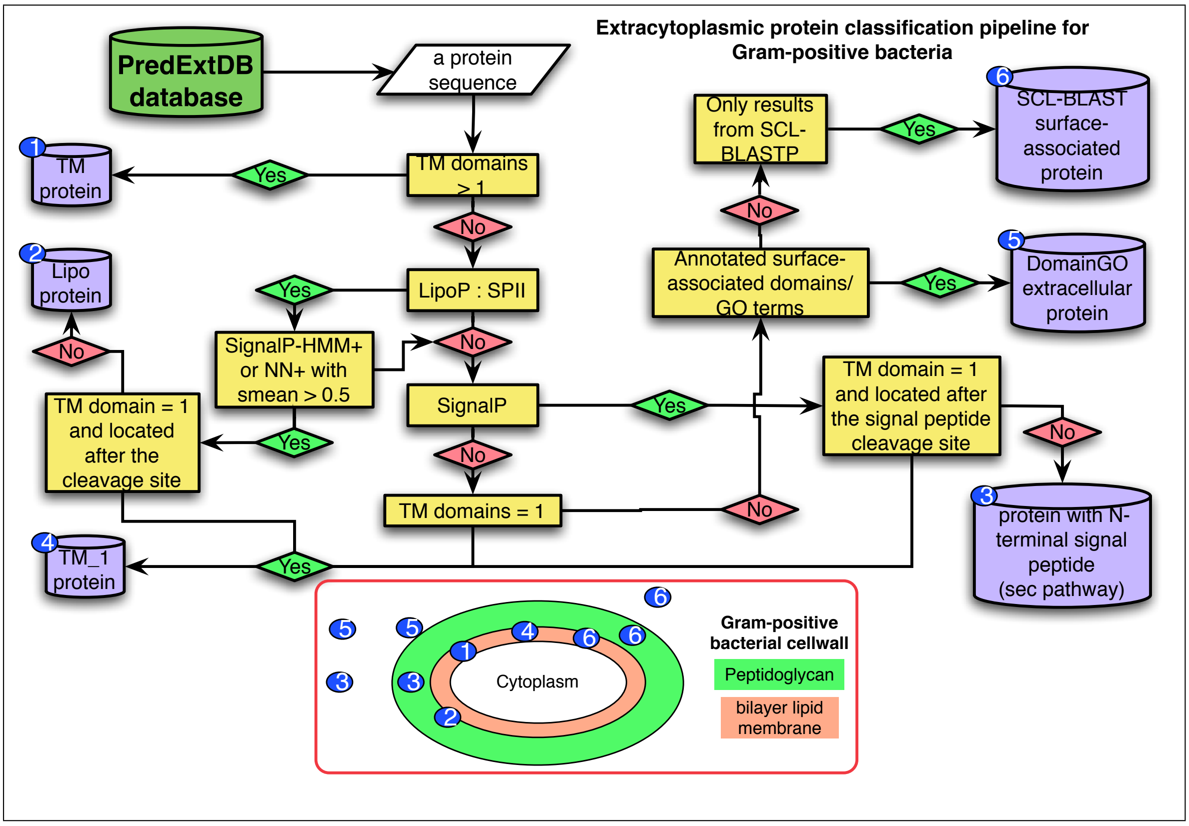

Extracytoplasmic protein classification pipeline for Gram-positive bacteria
PredExtDB database
a protein sequence
TM domains > 1
Yes
1 TM protein
No
LipoP : SPII
Yes
2 Lipo protein
SignalP-HMM+ or NN+ with smean > 0.5
No
Yes
TM domain = 1 and located after the cleavage site
SignalP
Yes
No
TM domains = 1
4 TM_1 protein
Yes
No
TM domain = 1 and located after the signal peptide cleavage site
No
3 protein with N-terminal signal peptide (sec pathway)
Annotated surface-associated domains/ GO terms
Yes
5 DomainGO extracellular protein
No
Only results from SCL-BLASTP
Yes
6 SCL-BLAST surface-associated protein
Cytoplasm
Gram-positive bacterial cellwall
Peptidoglycan
bilayer lipid membrane

Figure 4.4: Extracytoplasmic (ExCyt) protein classification pipeline for Gram-positive bacteria. Putative Excyt proteins were identified from TMHMM, SignalP, LipoP, SCL-BLASTP (search against a set of experimentally verified surface-associated proteins) and surface-associated protein domain or GO terms annotations. The set of predicted ExCyt proteins were retrieved from the analysis database, PredExtDB. A protein was systematically classified into one of six different classes: 1) Transmembrane (TM) protein if more than two helices were identified; 2) Lipoprotein if the SPII cleavage site was predicted by LipoP without a TM located after the cleavage site and with either a predicted N-terminal signal peptide by SignalP-HMM or SignalP-NN with a Smean score > 0.5; 3) Secreted protein via the Sec pathway if predicted positive by SignalP but no TM domain predicted, or a predicted TM region is located prior to the N-terminal signal peptide (SP) cleavage site; 4) Protein with a single TM domain if a TM domain was identified without being SignalP/Lipoprotein positive or a TM was located C-terminally to the cleavage site; 5) Protein with surface-associated domains/GO terms if annotated as such; 6) SCL-BLASTP protein if having a BLASTP hit with an e-value < 1E-9 by the SCL-BLASTP analysis. Each putative ExCyt protein can only be classified into one of the six defined classes as determined by the described pipeline. Data storage for each class is shown in purple. Yellow squares represent processing steps. Yes/No decisions made for each step throughout the classification pipeline are highlighted in green and red, respectively. Arrows show directions of the workflow.

Approach for classifying eukaryotic microbial extracytoplasmic proteins

Since microbial eukaryotic extracytoplasmic proteins have no lipoprotein signal peptides, the sequences were classified into five classes, disregarding the lipoprotein classification. No tools for predicting GPI-anchored proteins were included in this workflow for the reasons described in Section 3.3.2. The workflow assigns eukaryotic protein sequence into one of the following classes: multiple-

TM class, one-TM class, Sec-pathway class, surface-associated protein domains and GO terms class and homologs of verified extracytoplasmic membrane class (Figure 4.5). The eukaryotic-specific GPI-anchored proteins were anticipated to be classified into one of the extracytoplasmic class, either the Sec-pathway or TM classes.

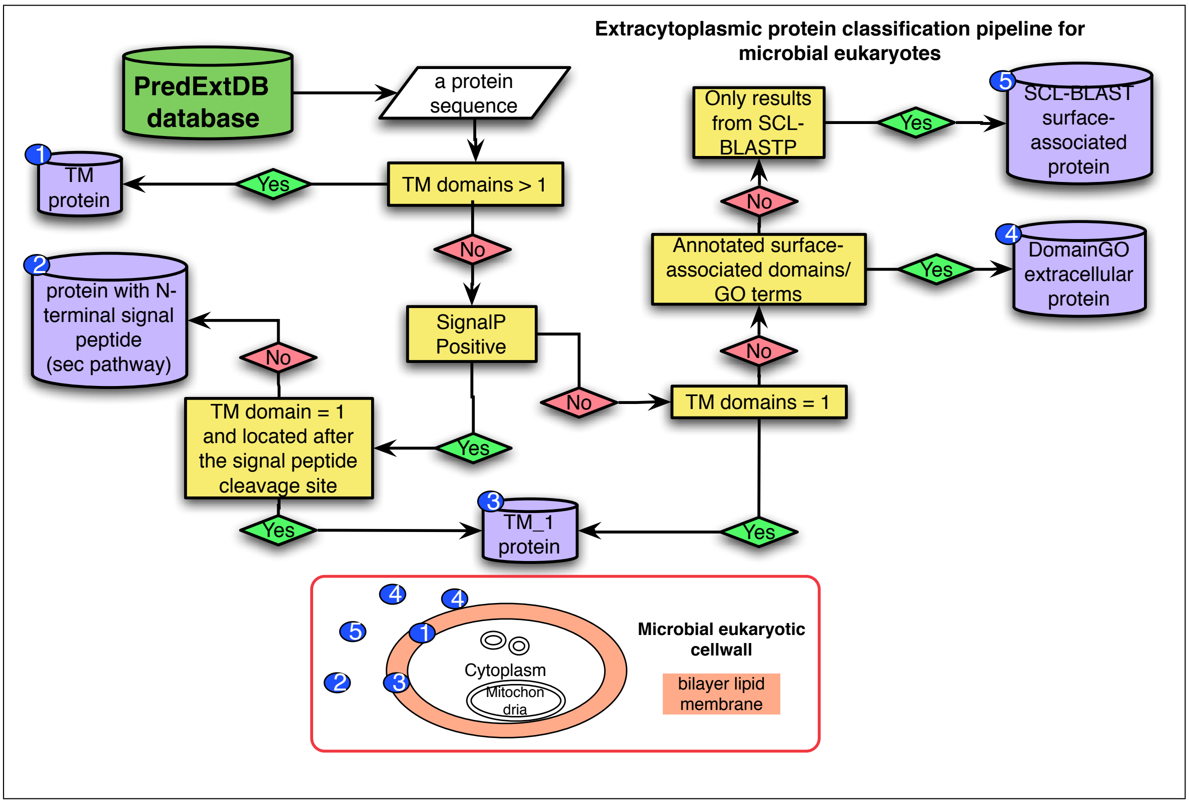

Extracytoplasmic protein classification pipeline for microbial eukaryotes
PredExtDB database
a protein sequence
TM domains > 1
Yes
1 TM protein
No
SignalP Positive
2 protein with N-terminal signal peptide (sec pathway)
No
TM domain = 1 and located after the signal peptide cleavage site
Yes
Yes
3 TM_1 protein
No
TM domains = 1
Yes
No
Annotated surface-associated domains/ GO terms
Yes
4 DomainGO extracellular protein
No
Only results from SCL-BLASTP
Yes
5 SCL-BLAST surface-associated protein
4
4
5
1
2
3
Cytoplasm
Mitochondria
Microbial eukaryotic cellwall
bilayer lipid membrane

Figure 4.5: Extracytoplasmic (ExCyt) protein classification pipeline for microbial eukaryotes. Putative ExCyt proteins resulting from TMHMM, SignalP, SCL-BLASTP (searches against a set of experimentally verified surface-associated proteins) and surface-associated protein domain/GO terms annotations. The set of predicted ExCyt proteins were retrieved from our analysis database. A protein was systematically classified into five different classes: 1) Transmembrane (TM) protein if more than two helices were identified; 2) Secreted protein via the Sec pathway if predicted positive by SignalP but no TM region was predicted, or a predicted TM domain was located prior to the N-terminal signal peptide (SP) cleavage site; 3) Protein with a single TM domain if a TM domain was identified without a positive SignalP result, or the TM was located C-terminally to the SP cleavage site; 4) Protein with surface-associated domains/GO terms if annotated as such; 5) SCL-BLAST Protein if a BLAST hit with an e-value < 1E-9 by SCL-BLASTP analysis. Each putative ExCyt protein can only be classified into one of the five defined classes as determined by the described pipeline. Data storage for each class is shown in purple. Yellow squares represent processing steps. Yes/No decisions made for each step throughout the classification pipeline are highlighted in green and red, respectively. Arrows show the direction of the workflow.

4.2.3 Performance evaluation of the extracytoplasmic protein identification workflow

To evaluate the performance of the universal extracytoplasmic protein identification workflow, results yielded from the workflow were cross-checked with a set of proteins whose subcellular localisation have been experimentally verified. A set of 12,896 verified protein sequences was obtained from ePSORT database (ePSORTdb) version 3 (accessed 28th March 2010) [Rey *et al.*, 2005]. This list contains proteins from archaea and bacteria. It is important to note that the list from ePSORTdb does not contain eukaryotic proteins. To perform a fair test, the sequences which were not included in this project (i.e., those proteins not available in the GPDB) were removed from the set of verified sequences, resulting in 9,265 remaining sequences. The resulting list was comprised of 6,745 cytoplasmic (cyto) proteins and 2,520 extracytoplasmic (non-cyto) proteins (see Table 4.1 for more details). Proteins were assigned as 'cyto' if they were shown to be only located in the cytoplasmic space, whereas the tag 'non-cyto' was assigned if a protein was exported from cytoplasmic space. The latter case included proteins which were experimentally verified to be translocated to the cytoplasmic membrane, periplasmic, outer membrane and extracellular spaces (secreted). To simplify the process of evaluation, the taxonomic groups described in this section were classified based on the organism group classification used in the ePSORTdb. ePSORTdb classifies prokaryotes into 5 groups: archaea, Gram-negative bacteria (Gm-), Gram-positive bacteria (Gm+), Gm- without outer membrane (Gm-/OM-), and Gm+ with outer membrane (Gm+/OM+). Notably, the nomenclature of the Gram-staining classification used in ePSORTdb is different from the nomenclature used by this study. In this project, Gm+ and Gm- were assigned to bacteria with respect to the bacterial taxonomic classification (see Figure 3.12), the cell surface structure was represented in the taxonomic classes (bacterial phylum), and the evolutionary relatedness of each phylum on the recent global phylogenetic tree was constructed based on universal protein families [Ciccarelli *et al.*, 2006]. The differences between the Gram-staining annotation between this project and the ePSORTdb can be seen. For example, Gm+/OM+ class in the ePSORTdb comprises some members of Deinococci phylum, e.g. *Deinococcus radiodurans* and *Deinococcus geothermalis* DSM 11300. These organisms were classified as Gm- by this project. In this project, the Tenericutes phylum was considered to be a Gm+, whereas ePSORTdb refers to this phylum as Gm-/OM-. Tenericutes was assigned the Gm+ class in this project because its bacterial members have single cell membrane and they are more closely related to the Firmicutes phylum whose members are known Gm+.

Performance evaluation metric

The metric used to perform the evaluation relied on four basic values — true positives (TP), false negatives (FN), false positives (FP) and true negatives (TN). To assess the performance of the workflow in term of identifying extracytoplasmic proteins, these statistics were calculated as shown in Table 4.2.

Positive predictive value was calculated as TP / (TP + FP), where as sensitivity (recall) was calcu-

Table 4.1: The number of proteins with an experimentally verified subcellular localisation obtained from ePSORTdb. These numbers were used as a baseline to assess the performance of the project's extracytoplasmic protein identification workflow. Organism groups noted in the table were obtained from the classification used in the ePSORTdb. CW = Cellwall, C = Cytoplasmic, CM = Cytoplasmic membrane, EC = Extracellular, OM = Outer membrane, P = Periplasmic, Gm- = Gram-negative bacteria, Gm+ = Gram-positive bacteria, OM- = no outer membrane.

Experimental localisation	Archaea	Gm-	Gm-/ OM-	Gm+	Gm+/ OM+	Total
CW	14	-	-	40	0	54
CW, EC	1	-	-	5	0	6
C	-	4,942	106	1,658	39	6,745
C, CM	-	39	0	35	0	74
CM	51	1,152	15	250	1	1,469
CM, CW	-	-	-	21	0	21
EC	9	177	0	73	0	259
OM	-	298	-	-	-	298
OM, EC	-	35	0	-		35
P	-	264	0	-	-	264
P, CM	-	33	0	-	-	33
P, OM	-	7	-	-	-	7
Total	75	6,947	121	2,082	40	9,265

Table 4.2: Basic values used to evaluate the performance of the extracytoplasmic protein identification workflow. Cyto = Cytoplasmic, Non-cyto = Non-cytoplasmic,TN = True positive, FN = False negative, TP = True positive, FP = False positive.

Actual localisation	**Predicted localisaion**	
	Cyto	Non-cyto
Cyto	TN	FP
Non-cyto	FN	TP

lated as TP / (TP + FN). Positive predictive value reflects the ability of the workflow to generate correct predictions. For example, a 95% positive predictive value would mean that for 100 predicted extracytoplasmic sequences, five are FPs or cytoplasmic. Sensitivity represents the ability of the workflow to identify all TPs or extracytoplasmic proteins. For example, 95% sensitivity indicates that for 100 actual extracytoplasmic sequences, five will be predicted as FNs or cytoplasmic proteins [Gardy and Brinkman, 2006].

4.3 Results

In this section, the outcomes of applying the workflows described in the previous section to the proteomes included in this study were presented. The performance of the workflows were evaluated by comparing the results to the set of proteins of known protein localisation. This section includes an application use case for the results generated from the workflows described in this chapter and chapter 3 to gain a greater understanding of the extracytoplasmic proteomes of the selected 24 *Bacillus* strains.

4.3.1 Comparison of the classification results to experimentally verified protein localisation

To evaluate the specificity and sensitivity of our protein subcellular identification approach, the results yielded from the workflow were compared with the experiment data of protein location derived from ePSORTdb (see Section 4.2.3). The experimental data set was used to measure the quality of the approaches used in this study for archaeal, Gram-positive and Gram-negative bacterial proteins.

The performance evaluation of the universal extracytoplasmic protein identification workflow

The performance evaluation of the universal extracytoplasmic protein identification workflow showed that it is possible to make a reliable prediction of extracytoplasmic proteins from the workflow described in this chapter. The performance was measured in comparison with the five currently available experimental data sets containing a total of 9,265 prokaryotic protein sequences (see methods Section 4.2.3). It is also important to note that the term ‘extracytoplasmic proteins’ used in this study are proteins located in any subcellular site, except the cytoplasmic space (see Figure 2.8). The performance of the workflow was calculated based on five experimental data sets in order to take into account the differences in the cell surface structures of distantly-related prokaryotes. The five data sets obtained from ePSORTdb were archaeal, Gm+, Gm-, Gm+/OM+ and Gm-/OM-. Positive predictive value and sensitivity of the workflow were computed for these five organism groups. The overall positive predictive value reached 100%, 90.8%, 95.2%, and 87.5% for each group respectively, except for Gm+/OM+ data set which had only 25% positive predictive value. The low positive predictive value of Gm+/OM+ prediction might be due to the very low number of extracytoplasmic proteins in the experimental data set (only one protein was verified to be on the cytoplasmic membrane; see Table 4.1). The sensitivities of the workflow were: 90.7%, 86.6%, 88.7%, 93.3% and 100%, respectively for each data set. Further details are shown in Table 4.3. Based on the performance evaluation using experimentally verified protein data sets, it was difficult to evaluate the performance of the workflow on the archaeal protein data set as there were only a small numbers (75) of archaeal proteins with experimentally verified locations.

Table 4.3: Performance of the project's workflow for the identification of extracytoplasmic proteins. TN = True positive, FN = False negative, TP = True positive, FP = False positive, Gm- = Gram-negative bacteria, Gm+ = Gram-positive bacteria, OM- = no outer membrane.

Organism group	TP	FP	TN	FN	Positive predictive value	Sensitivity
Archaea	68	0	0	7	100.00%	90.67%
Gm-	1,779	89	4,853	226	95.24%	88.73%
Gm-/OM-	14	2	104	1	87.50%	93.33%
Gm+	367	37	1,621	57	90.84%	86.56%
Gm+/OM+	1	3	36	0	25.00%	100.00%

***Bacillus subtilis* lipoprotein prediction**

To evaluate the performance of the workflow for the identification of Gram-positive lipoproteins, the results from the Gram-positive workflow were cross-checked with a list of putative *B. subtilis* lipoproteins proposed by Tjalsma et al [Tjalsma *et al.*, 2000]. The list of lipoproteins from Tjalsma *et al.*'s study combined experimentally-verified lipoproteins and a list of putative lipoproteins that were identified by manually checking for regions likely to be lipoprotein signal peptides. In our workflow, proteins were identified as Gram-positive lipoproteins if their sequence had positive predictions from both LipoP and SignalP (see Section 4.2.2).

Eighty-six out of 114 *B. subtilis* lipoproteins identified in the paper [Tjalsma *et al.*, 2000] were classified as lipoproteins by the workflow developed in this project. The workflow recognised eight more lipoproteins (YusW, Yscb, yfKR, Med, yddJ, yloI, yybP, YlbC) that were not listed as putative lipoproteins by Tjalsma *et al.* Ten lipoproteins (CtaC, SpoIIIJ, QoxA, YdiK, YhaR, YkoH, YqJG, YtrF, YwnJ, YybM) from *B. subtilis* predicted by Tjalsma *et al.* containing multiple alpha-helix membrane regions were assigned to the 'TM' class by the workflow. These proteins in the 'TM' class were therefore regarded as putative transmembrane proteins with multiple alpha-helical segments. It is noteworthy that only the first two proteins were also predicted to have an SPII cleavage site by LipoP. However, if these results from both sources were true, this discrepancy may suggest that these proteins possess more than one membrane-anchoring feature.

The YmzC protein was assigned into the one-TM class because it was predicted to have an N-terminal hydrophobic region of 21 amino acids by TMHMM (with predicted topology : i13-34o) and negative predictions by LipoP and SignalP with Gram-positive option selected. This protein was identified as a putative lipoprotein by Tjalsma *et al.* The protein also contains the twin arginine motif at the N-domain of the signal peptide, suggesting the export of the protein via the Tat pathway rather than the classical Sec pathway [Tjalsma *et al.*, 2000][Cristóbal *et al.*, 1999].

Performance of the comprehensive extracytoplasmic protein classification workflow

The results of the comprehensive extracytoplasmic protein classification workflow were compared to the localisation of protein sequences from the same experimental data sets described in the previous section (Section 4.3.1). The mapping result is shown in Table 4.4. Roughly 79% (1264/1597) of proteins that were experimentally proven to be localised on prokaryotic cytoplasmic membranes (CM) were classified as transmembrane proteins (multiple-TM or one-TM classes) by the workflow. Approximately 10.6% (169) of the verified CM proteins were systematically grouped into other extracytoplasmic protein classes including Sec-pathway, lipoprotein and proteins with known surface-associated domains. The remaining 10.4% (164) were not predicted as putative extracytoplasmic by the workflow.

Notably, the workflow did not incorporate any tool specifically intended for identifying Gram-negative bacterial outer membrane (OM) beta-barrel proteins, so it is worthwhile to examine the outcome. The detection of OM proteins relied on the SCL-BLASTP search and protein signatures and domains of known OM proteins (see Table 3.1). Approximately, 97.4% (331/340) of the verified OM proteins were predicted as putative extracytoplasmic proteins by the workflow. Most of them were classified as Sec-pathway proteins. Some OM proteins were predicted to have lipoprotein signal peptides, and a few were predicted to have alpha-helix transmembrane regions. The rest were filtered by the workflow as extracytoplasmic proteins with known surface-associated domains or GO terms (see Table 4.4). From the results, it was noticeable that most of the OM proteins were exported via the Sec pathway due to the presence of Sec signal peptides. This characteristic of the beta-barrel outer membrane proteins has already been observed by other studies [Bagos *et al.*, 2004a][Bagos *et al.*, 2004b]. The presence of alpha-helix regions in the OM proteins might be due to the fact that some OM proteins are known to possess alpha-helical hydrophobic regions [Noppa *et al.*, 2001][Bunikis *et al.*, 1995].

The aim of the project was to identify extracytoplasmic proteins regardless of where they are located. The approaches used and workflows developed in this chapter cover 88.5% (2229/2520) of the extracytoplasmic proteins from verified archaeal and bacterial proteins with various cell surface structures. The remaining 11.5% (291) of the verified extracytoplasmic proteins that were not identified as extracytoplasmic proteins by the workflow were investigated manually. It appears that 227 of these proteins were Gram-negative bacterial proteins of which 87, 72 and 26 are verified as cytoplasmic membrane, extracellular and periplasmic proteins, respectively. Five were verified OM proteins and 33 were identified in both cytoplasmic and CM, whereas four were presented as either OM or extracellular. This finding suggests that several Gram-negative bacterial extracytoplasmic proteins are not exported via the classical Sec pathway nor do they have any alpha-helix transmembrane segments or other detectable surface-associated features. This might be due to the variation of Gram-negative bacterial secretory machinery. For example, many virulence-related proteins secreted through type III secretory system and other non-classical secretory pathways do not have any well-conserved regions nor recognisable targeting signal sequences [Samudrala *et al.*, 2009][Arnold *et al.*, 2009].

Furthermore, the performance of the workflow in the classification of transmembrane proteins ('TM' class) ranged from 81-96% positive predictive value and 84-92% sensitivity for different prokaryote groups (see Table 4.5 for more details). It is important to note that the transmembrane proteins class here were defined by TMHMM, detecting the presence of alpha helices. These proteins are mostly located on the cytoplasmic membrane (inner membrane) of the Gram-negative bacteria. For proteins localised on the Gram-negative outer membrane, they are typically presented with beta-barrel or particular motifs (see Table 3.1).

For the proteins classified as putative secreted and surface-anchoring proteins ('Sec' class), 12% (96/786) of the predicted Sec-class proteins were experimentally verified as cytoplasmic proteins. The Sec-class included sequences with SignalP predicted positives with no predicted alpha-helix membranes and were not predicted as lipoproteins. The positive predictive value of the workflow for the identification of secreted and surface proteins were 61%, 73% and 80% for Gm+, Gm- and archaea, respectively. The sensitivity of these groups of organisms were 74%, 82% and 46%, respectively (see Table 4.6).

Table 4.4: Results of the classification of extracytoplasmic proteins ('ext') using the project's comprehensive workflow in relation to the data set of the experimentally verified protein sublocalisation. extblst = 'ext' predicted by SCL-BLASTP, extdom = 'ext' predicted by having surface-associated protein domains or gene ontology terms, lipo = 'ext' having signal peptidase II cleavage site, Sec = 'ext' having signal peptidase I cleavage site, TM = 'ext' having at least one putative alpha-helix transmembrane region(s), Gm- = Gram-negative bacteria, Gm+ = Gram-positive bacteria, OM- = no outer membrane, CW = Cellwall, C = Cytoplasmic, CM = Cytoplasmic membrane, EC = Extracellular, OM = Outer membrane, P = Periplasmic, OM- = no outer membrane.

	Predicted extracytoplasmic classes					
Experimental localisation	**extblst**	**extdom**	**lipo**	**Sec**	**TM**	**Total**
Archaea						
CM	-	-	-	4	43	47
CW	-	-	-	6	8	14
CW, EC	-	-	-	1	-	1
EC	-	-	-	4	2	6
Gm-						
C	-	3	5	61	20	89
C, CM	-	-	-	5	3	8
CM	-	1	14	78	972	1,065
P	-	-	7	226	5	238
P, CM	-	4	2	10	16	32
P, OM	-	-	6	-	1	7
EC	-	9	1	81	14	105
OM	1	6	60	223	3	293
OM, EC	-	1	1	26	3	31
Gm-/OM-						
C	-	-	-	2	-	2
CM	-	1	1	1	11	14
Gm+						
C	-	2	-	30	5	37
C, CM	-	-	-	5	22	27
CM	-	-	10	20	191	221
CM, CW	-	-	2	11	5	18
CW	-	4	1	24	8	37
CW, EC	-	-	-	5	-	5
EC	-	-	-	56	3	59
Gm+/OM+						
C	-	-	-	3	-	3
CM	-	-	-	-	1	1
Total	1	31	110	882	1,336	2,360

Table 4.5: Performance of the project's extracytoplasmic classification workflow for the identification of transmembrane proteins. The results of the classified transmembrane sequences were compared to the experimentally-verified cytoplasmic membrane (excluding Gram-negative outer membrane proteins). Gm- = Gram-negative bacteria (excluding the Gm-/OM- data set), Gm+ = Gram-positive bacteria (excluding the Gm+/OM+ data set), TP = true positive, FP = false positive, FN = false negative.

Organism group	TP	FP	FN	Positive predictive value	Sensitivity
Archaea	43	10	4	81.13%	91.49%
Gm-	992	45	119	95.66%	89.26%
Gm+	218	16	40	93.16%	84.50%

Table 4.6: Performance of the project's extracytoplasmic classification workflow for the identification of secretome and surface proteins. The results of the classified secreted and surface protein sequences were compared to the experimentally-verified cell wall, extracellular and Gram-negative outer membrane proteins.

Organism group	TP	FP	FN	Positive predictive value	Sensitivity
Archaea	11	4	13	73.33%	45.83%
Gm-	664	167	150	79.90%	81.57%
Gm+	103	67	36	60.59%	74.10%

4.3.2 Large-scale extracellular protein classification

The workflows were applied to 3,021,490 protein sequences in the GenomePool database to identify putative extracytoplasmic sequences and their potential specific extracytoplasmic localisations. Figure 4.6 summarises the proportion of predicted extracytoplasmic proteins across different groups of microorganisms. Table 4.7 provides a summary of the organism types and classes, referring to the subcellular locations of the protein sequences classified using the workflow described in this chapter. Based on the proteomes included in this study, the fractions of putative extracytoplasmic proteins across the four groups of microorganisms were estimated to be 24.6%, 25.9%, 31%, and 34.6% for microbial eukaryotes, archaea, Gram-positive bacteria and Gram-negative bacteria, respectively.

The 'TM' class accounted for the largest fraction of the putative extracytoplasmic proteins in all four organism groups. The fraction of the transmembrane proteins ranged from 15.3% in the microbial eukaryote group to 20.8% in the Gram-positive bacterial group (see Figure 4.7). The percentages of the fractions presented here were computed in proportion to all protein sequences in each group of organisms. The results indicated that Gram-negative bacteria and archaea carry a relatively similar proportion of alpha-helix transmembrane proteins: 19% and 18.7% of the proteome data set, respectively. These transmembrane proteins are typically translocated from cytoplasm to the cytoplasmic membrane via the universal Sec pathway. The proteins exported via the classical Sec pathway were classified into the 'Sec' class which included several extracytoplasmic proteins such as cell surface-anchoring, Gram-negative outer membrane, and periplasmic proteins.

Table 4.7: Summary of protein sequences assigned to different classes by the extracytoplasmic classification workflow. The workflow was applied to all protein sequences deposited in the GenomePool database. The results were shown in relation to organism groups depending on the major cell surface structures. The Gram-positive group includes members of bacterial phyla Actinobacteria, Firmicutes and Tenericutes. Other bacterial phyla are considered to belong to the Gram-negative group. The number of sequences were counted based on the sequence classes assigned by the project workflow. extprot = putative extracellular cytoplasmic proteins, extblast = 'extprot' predicted by SCL-BLASTP, extdom = 'extprot' predicted by having surface-associated protein domains or gene ontology terms, lipo = 'extprot' having signal peptidase II cleavage site, Sec = 'extprot' having signal peptidase I cleavage site, TM = 'extprot' having at least one putative alpha-helix transmembrane region(s), Gm- = Gram-negative bacteria, Gm+ = Gram-positive bacteria.

Organism group	Total proteins	TM	Sec	lipo	extdom	extblast	Total extprot
Gm+	693,402	144,290	55,092	13,911	1,648	14	214,955
Gm-	1,922,673	365,395	249,166	44,630	5,903	100	665,194
Eukaryotic	272,389	41,743	24,510	-	866	2	67,121
Archaea	133,026	24,826	7,975	1,527	170	1	34,499
Total	3,021,490	576,254	336,743	60,068	8,587	117	981,769

The 'Sec' class contained approximately 6%, 8%, 9% and 13% of the proteome of archaea, Gram-positive bacteria, microbial eukaryotes, and Gram-negative bacteria, respectively (see Figure 4.7). Notably, the Gram-negative bacterial group carried a slightly higher fraction of Sec-signal proteins without alpha-helical transmembrane regions than those from other microorganism groups. These proteins could contribute in the Type II or V protein secretion systems which are found predominantly in the Gram-negative bacteria (see Section 2.3.3). It is known that Gram-negative bacterial proteins located on outer membrane often contain beta-barrel transmembrane regions. These outer-membrane proteins are often exported from cytoplasm across the inner membrane by the Sec pathway before forming a beta-barrel sheet and inserting themselves into the outer membrane [Wimley, 2003][Cullen, 2004]. On the other hand, predicted 'Sec-pathway' sequences for Gram-positive bacteria could be secreted to the extracellular space or located or anchored on a peptidoglycan layer. In the case of eukaryotic microbial proteins predicted as 'Sec-pathway', many of these proteins could be either extracellular proteins or might instead be retained in cytoplasmic organelles such as endoplasmic reticulum or golgi [Dyrlovbendtsen, 2004]. Likewise, predicted 'transmembrane' eukaryotic proteins would also include sequences located on the membrane of the organelles in eukaryotic cells as well as the cytoplasmic cell membrane. Lipoproteins were predicted in very narrow ranges of 1.2%-2.3% across archaeal and bacterial proteomes. Nearly the same proportions of putative lipoproteins were observed in Gram-positive and Gram-negative bacterial proteomes (2% and 2.3%, respectively).

Moreover, based on the extracytoplasmic protein prediction workflows, there was a strong positive correlation between the proteome size and the size of alpha helical transmembrane proteins ($R^2 \geq 0.89$) among all group of microorganisms (see Figure 4.8). Likewise, the positive correla-

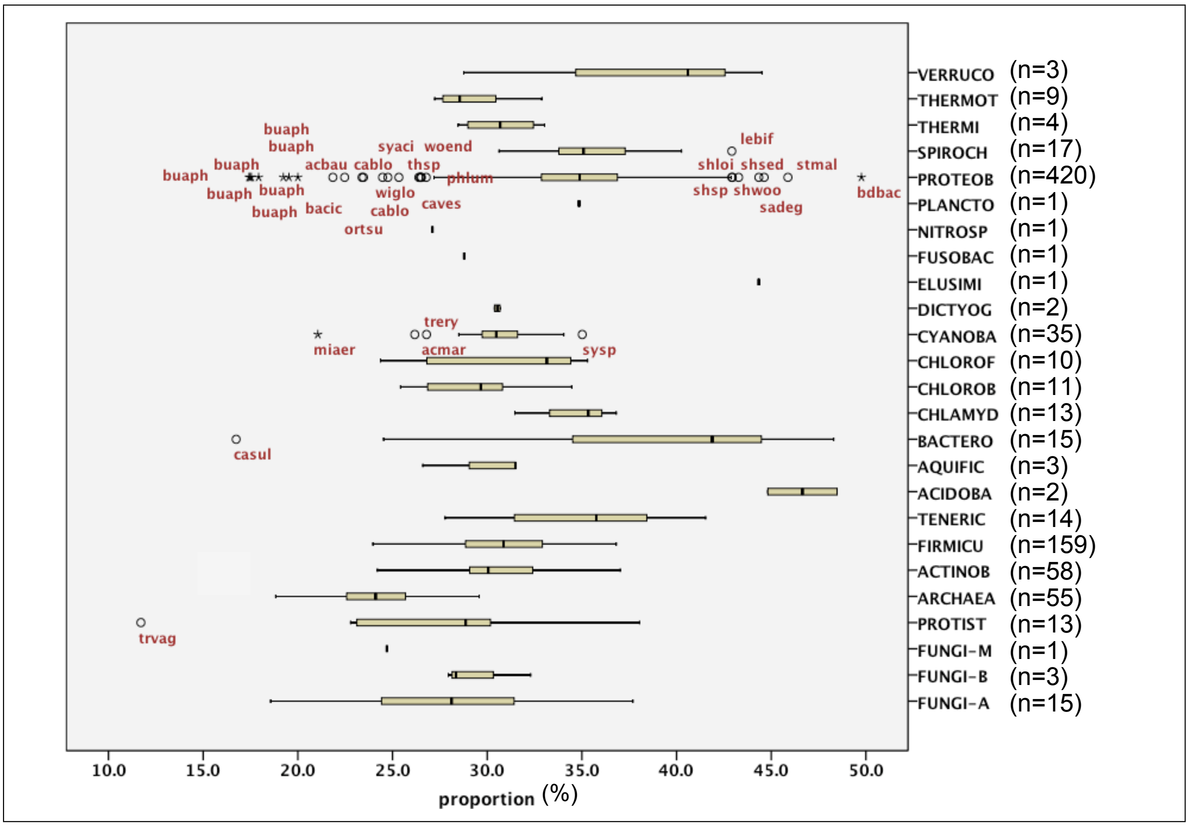


Figure 4.6: A boxplot displaying proportions of predicted extracytoplasmic proteins in different phyla. The vertical axis contains the boxplots for different taxonomic groups. The horizontal axis represents the average proportion of the predicted extracytoplasmic proteins across the total number of protein sequences in a given taxonomic group. A proportion was calculated into percentage for each organism. Extracytoplasmic proteins were predicted by a combination of results obtained from several bioinformatics tools including TMHMM, SignalP, LipoP, InterProScan, and BLASTP as described in Section 4.2.1. The resulting proportions were plotted with respect to organism taxonomic groups. The lower and upper edges of each box indicate the 25th and 75th percentiles, respectively, of the values found in a particular taxonomic group. The vertical line in each box indicates the median value of the data. The ends of the horizontal lines (whiskers) indicate the minimum and maximum data values. The whisker extends to a maximum of one quarter of the data unless outliers are present. Outliers are shown by open circles. Asterisks denote extreme outliers. n=number of taxa. buaph= *Buchnera aphidicola* str. Tuc7 (*Acyrthosiphon pisum*), *Buchnera aphidicola* str. Sg (*Schizaphis graminum*), *Buchnera aphidicola* str. Cc (*Cinara cedri*), *Buchnera aphidicola* str. Bp (*Baizongia pistaciae*), *Buchnera aphidicola* str. APS (*Acyrthosiphon pisum*), *Buchnera aphidicola* str. 5A (*Acyrthosiphon pisum*), *Buchnera aphidicola* (*Cinara cedri*), acbau = *Acinetobacter baumannii* ATCC 17978, bacic= *Baumannia cicadellinicola* str. Hc (*Homalodisca coagulata*), cablo= *Candidatus Blochmannia* floridanus , *Candidatus Blochmannia* pennsylvanicus str. BPEN, wiglo= *Wigglesworthia glossinidia* endosymbiont of *Glossina brevipalpis*, ortsu= *Orientia tsutsugamushi* str. Ikeda, syaci = *Syntrophus aciditrophicus* SB, thsp= Thauera sp. MZ1T, woend= *Wolbachia endosymbiont* of *Drosophila melanogaster* , phlum= *Photorhabdus luminescens* subsp. laumondii TTO1, caves= *Candidatus Vesicomyosocius* okutanii HA (C*andidatus Vesicomyosocius* okutanii str. HA) ,lebif= *Leptospira biflexa* serovar Patoc strain 'Patoc 1 (Ames)', shloi= *Shewanella loihica* PV-4, shsed= *Shewanella sediminis* HAW-EB3, shsp= *Shewanella sp.* MR-4, shwoo= *Shewanella woodyi* ATCC 51908 , sadeg= *Saccharophagus degradans* 2-40, stmal= *Stenotrophomonas maltophilia* R551-3, bdbac= *Bdellovibrio bacteriovorus* HD100, miaer= *Microcystis aeruginosa* NIES-843, trery= *Trichodesmium erythraeum* IMS101, acmar= *Acaryochloris marina* MBIC11017, sysp= *Synechococcus sp.* WH 8102, casul= *Candidatus Sulcia muelleri* GWSS, trvag =*Trichomonas vaginalis* G3.

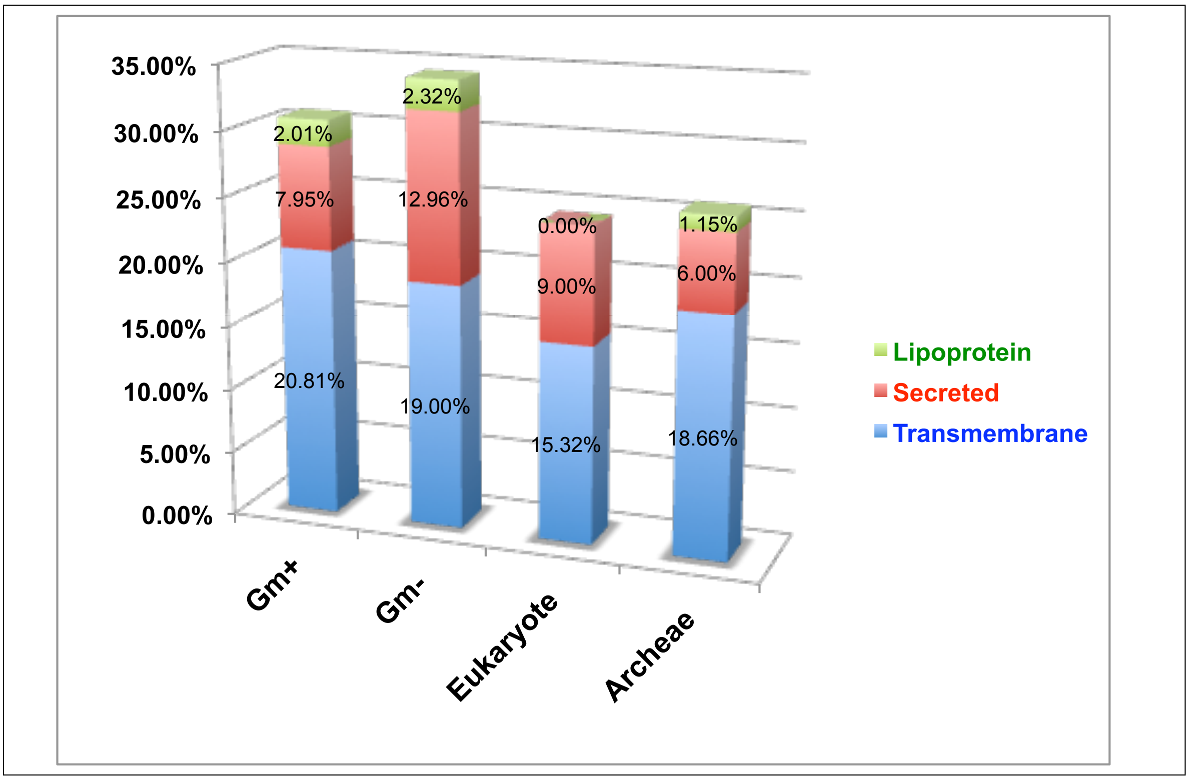


Figure 4.7: Proportions of extracytoplasmic proteins among all of the protein sequences across different types of organisms. The proportions are shown as percentages of the predicted extracytoplasmic proteins in a given class across the total number of protein sequences in a particular organism group. The Gram-positive group (Gm+) included members of bacterial phyla Actinobacteria, Firmicutes and Tenericutes. Other bacterial phyla were considered as Gram-negative group (Gm-).

tion was observed between the proteome size and the size of ‘Sec-pathway’ proteins ($R^2 \geq 0.73$). The results suggest that the larger the number of sequences in a proteome, the greater the number of cytoplasmic membrane proteins and proteins carrying Sec-signal peptides. The weaker positive correlations were observed between the proportion of predicted lipoproteins and proteome size ($0.60 \leq R^2 \leq 0.63$) across archaea and Gram-positive bacteria. The least positive correlation ($R^2 = 0.31$) was found between the fraction of lipoproteins and proteome size of the Gram-negative data set. For the eukaryotic data set, *Trichomonas vaginalis G3*’s proteome was excluded from the plot in Figure 4.8 to remove an extreme outlier as the genome encodes approximately 59,518 protein-coding genes. The size of the *T. vaginalis G3* proteome is extremely large compared to those from the other microbial eukaryotes (ranges from 403 to 13,331 protein sequences) in this study. Nonetheless, the Trichomonas’ proteome appeared to contain the smallest fraction of extracytoplasmic proteins (11.7%) according to the results from the workflow developed in this project (see Figure 4.6).

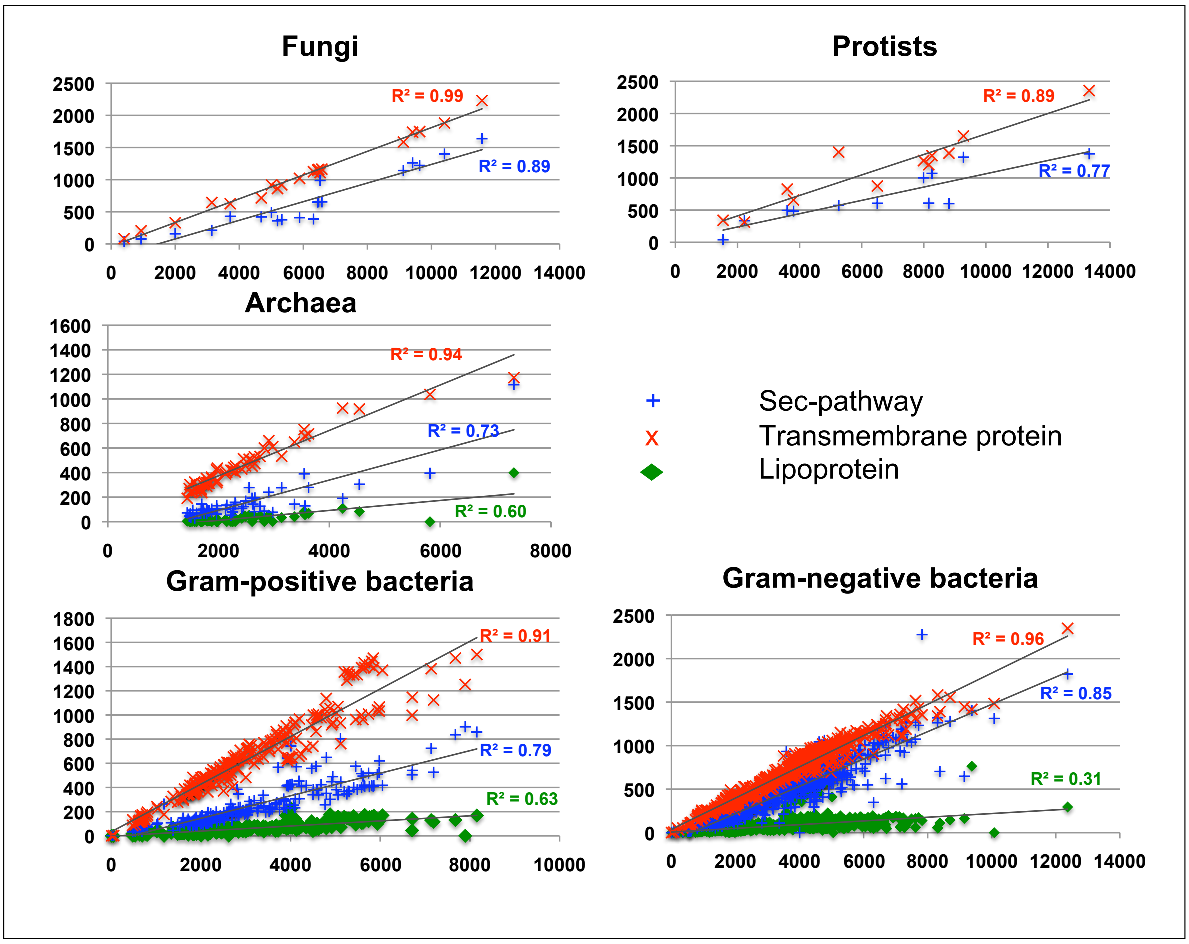


Figure 4.8: The correlation between the numbers of extracytoplasmic proteins and the total number of protein sequences is shown for five groups of microorganisms. The predicted extracytoplasmic proteins were classified into lipoprotein, alpha-helix transmembrane proteins and Sec-pathway proteins. Note that Sec-pathway class included proteins that possess Sec-signal peptides and have not yet been classified into the lipoprotein or the transmembrane class. The X-axis shows the total number of protein sequences in each proteome. The Y-axis denotes number of sequences predicted as putative extracytoplasmic proteins. This plot excludes an extreme outlier i.e. *Trichomonas vaginalis G3*'s proteome data set.

4.3.3 Extracytoplasmic proteome prediction of *Bacillus spp.* using the Microbase workflows

In this section, an application use case is demonstrated that makes use of the results generated from the workflow described in the previous sections. An analysis of 24 proteomes of *Bacillus spp.* was performed that covered all the completed *Bacillus* genomes available at the time of study.

Bacillus spp. is a rod-shaped, spore-forming Gram-positive bacteria belonging to the phyla Firmicutes. The *Bacillus* members appear as both being free-living and being associated with hosts. They also exhibit differences in terms of host range and virulence. Some *Bacillus* species are pathogenic to insects or vertebrates. Phylogenetic analysis of the 16sRNA demonstrated a close relationship among some *Bacillus* species that are considered as host-associated and pathogenic in some insects and mammals. This group of genetically closely-related Bacillus species, known as the *Bacillus cereus* group, includes *B. cereus*, *B. anthracis*, *B. thuringiensis*, *B. weihenstephanensis*, *B. cytotoxicus* and *B. mycoides* [Kolstø *et al.*, 2009] (see Figure 4.9). The first three species are known to be pathogenic to mammals or insects, whereas the last two species and the members of non-cereus group are generally regarded as non-pathogenic soil bacteria (see Figure 4.9 for the list of non-cereus group's members). The term 'cereus group' is used to refer to this closely-related *Bacillus* species throughout this section.

The variation within the *Bacillus* species in terms of their ecological niches and symbiosis raises several interesting questions. For example, what is the diversity of the extracytoplasmic proteomes within the *Bacillus* species and how were the proteomes influenced by different ecological and evolutionary forces. In particular, what are the features that facilitate the cereus group's members in their interactions with hosts in comparison to the free-living non-cereus species. The availability of the genomes and the corresponding protein-coding gene sequences of the *Bacillus* species of both groups stimulated interest in the comparisons of their protein contents.

Genome sequences used for the analysis of the *Bacillus*' extracytoplasmic proteins

An analysis was performed on 57 *Bacillus* complete chromosomal and plasmid genomes from 24 *Bacillus* strains corresponding to 125,564 protein sequences being analysed in the workflow. The RefSeq genome data files were downloaded in GenBank (.gbk) format from the RefSeq database (on 27 July 2009). The 24 *Bacillus* taxa with complete genome sequences that were included in the analysis are shown in Table 4.8.

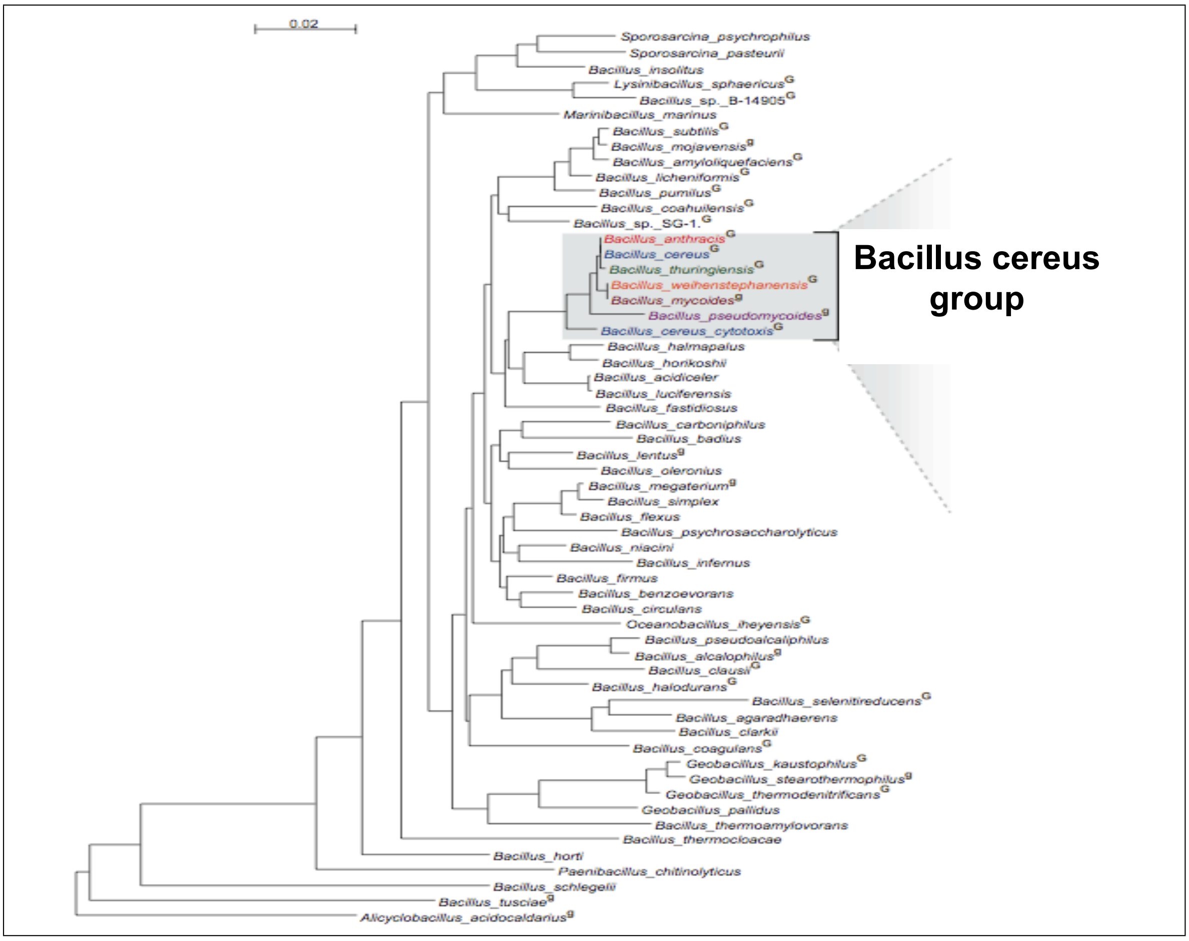

0.02
Sporosarcina_psychrophilus
Sporosarcina_pasteurii
Bacillus_insolitus
Lysinibacillus_sphaericus^{G}
Bacillus_sp._B-14905^{G}
Marinibacillus_marinus
Bacillus_subtilis^{G}
Bacillus_mojavensis^{g}
Bacillus_amyloliquefaciens^{G}
Bacillus_licheniformis^{G}
Bacillus_pumilus^{G}
Bacillus_coahuilensis^{G}
Bacillus_sp._SG-1.^{G}
Bacillus_anthracis^{G}
Bacillus_cereus^{G}
Bacillus_thuringiensis^{G}
Bacillus_weihenstephanensis^{G}
Bacillus_mycoides^{g}
Bacillus_pseudomycoides^{g}
Bacillus_cereus_cytotoxis^{G}
Bacillus cereus group
Bacillus_halmapalus
Bacillus_horikoshii
Bacillus_acidiceler
Bacillus_luciferensis
Bacillus_fastidiosus
Bacillus_carboniphilus
Bacillus_badius
Bacillus_lentus^{g}
Bacillus_oleronius
Bacillus_megaterium^{g}
Bacillus_simplex
Bacillus_flexus
Bacillus_psychrosaccharolyticus
Bacillus_niacini
Bacillus_infernus
Bacillus_firmus
Bacillus_benzoevorans
Bacillus_circulans
Oceanobacillus_iheyensis^{G}
Bacillus_pseudoalcaliphilus
Bacillus_alcalophilus^{g}
Bacillus_clausii^{G}
Bacillus_halodurans^{G}
Bacillus_selenitireducens^{G}
Bacillus_agaradhaerens
Bacillus_clarkii
Bacillus_coagulans^{G}
Geobacillus_kaustophilus^{G}
Geobacillus_stearothermophilus^{g}
Geobacillus_thermodenitrificans^{G}
Geobacillus_pallidus
Bacillus_thermoamylovorans
Bacillus_thermocloacae
Bacillus_horti
Paenibacillus_chitinolyticus
Bacillus_schlegelii
Bacillus_tusciae^{g}
Alicyclobacillus_acidocaldarius^{g}

Figure 4.9: Phylogenetic relationships among Bacilli members. Relationships among 57 *Bacillus* species based on 16S ribosomal DNA (rDNA) sequences. *Alicyclobacillus acidocaldarius* was used as an outgroup to root the tree. (Source: Klost et al. 2009 [Kolstø *et al.*, 2009]; more details of the phylogenetic tree construction can be found in the source paper)

Table 4.8: List of organisms and genomes used for the *Bacillus* proteome analysis. The RefSeq genome accession numbers are indicated. *Bacillus* genome information were derived from GOLD database. Asterisks indicate pathogenic strains.

Species	Refseq accession	Habitat	Isolation site	Phenotype
Cereus group				
B. anthracis str. A0248*	NC_012659, NC_012656, NC_012655		Human	Cause anthrax
B. anthracis str. Ames	NC_003997	Soil	-	Non-pathogen
B. anthracis str. ‘Ames Ancestor’*	NC_007530, NC_007322, NC_007323	Soil	-	Cause anthrax
B. anthracis str. CDC 684*	NC_012581, NC_012579, NC_012577	Soil	-	Cause anthrax
B. anthracis str. Sterne	NC_005945	Soil		Non-pathogen
B. cereus 03BB102*	NC_012472, NC_012473	Host, Soil	Blood of an infected human	Cause pneumonia
B. cereus AH187*	NC_011658, NC_011654, NC_011655, NC_011657, NC_011656	Host, Soil	Vomit of an infected human	Cause food poisoning
B. cereus AH820*	NC_011773, NC_011771, NC_011777, NC_011776	Soil	Periodontal pocket of a patient with marginal periodontitis	Cause food poisoning
B. cereus ATCC 10987*	NC_003909, NC_005707	Soil, Dairy isolate	Cheese spoilage	Cause food poisoning
B. cereus ATCC 14579*	NC_004722, NC_004721	Soil	-	Cause food poisoning
B. cereus B4264*	NC_011725	Host, Soil	Blood and the pleural fluid of an indefected human	Cause pneumonia
B. cereus E33L*	NC_006274, NC_007103, NC_007105, NC_007104, NC_007106, NC_007107	Soil	Swab of a zebra carcass	Cause food poisoning
B. cereus G9842*	NC_011772, NC_011774, NC_011775	Soil	Stool samples from an infected human	Cause food poisoning
B. cereus Q1	NC_011969, NC_011973, NC_011971	Soil, Oil fields	Deep-subsurface oil reservoir	Non-pathogen
B. cytotoxicus NVH 391-98	NC_009674, NC_009673	Soil	-	Non-pathogen
B. thuringiensis serovar konkukian str. 97-27*	NC_005957, NC_006578	Host, Soil	Severe human tissue necrosis	Cause sotto disease
B. thuringiensis str. Al Hakam*	NC_008600, NC_008598	Host, Soil	Severe human tissue necrosis	Cause sotto disease
B. weihenstephanensis KBAB4*	NC_010184, NC_010180, NC_010181, NC_010182, NC_010183	Soil	-	Non-pathogen
Non-cereus group				
B. amyloliquefaciens FZB42	NC_009725	Rhizosphere-colonizing, Soil	Soil	Non-pathogen
B. clausii KSM-K16	NC_006582,	Soil	-	Alkalitolerant
B. halodurans C-125	NC_002570,	Soil, Fresh water	-	Alkalophile
B. licheniformis ATCC 14580 (DSM 13)*	NC_006322, NC_006270	Soil	-	Cause food poisoning
B. pumilus SAFR-032	NC_009848	Soil	Spacecraft Assembly Facility at NASA Jet Propulsion Laboratory	Biomass degrader
B. subtilis subsp. subtilis str. 168	NC_000964	Soil	X-ray irradiated strain	Non-pathogen

General observations

The proportions of the extracytoplasmic proteomes of 24 *Bacillus* subspecies were 34.2%±5% (see Table 4.9 and Figure 4.11). The smallest extracellular proteome (31.6%) was *B. halodurans* C-125, whereas the largest (36.5%) was *B. thuringiensis* serovar konkukian str. 97-27. The proteome size of the cereus group's species was significantly larger than the non-cereus members (p-value < 0.05). The increase in extracytoplasmic proteome size increases with increase in proteome size with a strong positive correlation ($R^2 = 0.98$) (see Figure 4.10). The strong correlation between the extracytoplasmic proteome size and the proteome size from this study corroborated the finding of Gomi *et al.* [Gomi *et al.*, 2005a] whose secretome analysis observed the same pattern in Gram-positive bacteria. Gomi *et al.* predicted secreted and transmembrane proteins using a prediction tool called SOSUI and SOSUIsignal. These predictors rely purely on the physical properties of amino acid sequences and statistical analysis [Hirokawa *et al.*, 1998][Gomi *et al.*, 2005b]. However, this result contradicts the observations made by Song *et al.* [Song *et al.*, 2009] who found no correlation between the size of Gram-positive bacterial proteome and secretome from their study. However, Song *et al.* employed a different application, ExProt, to establish protein sets of belonging to the secretome. ExProt identifies the secretome by the presence of a bacterial lipoprotein motif (PS00013) and a signal peptide cleavage site for SPI through a amino acid position neural network and weight matrix algorithms [Saleh *et al.*, 2001]. This prediction tool excluded the proteins with a helix transmembrane domain and other known extracellular-associated domains which are included in our study. It was notable that all these assumptions, regarding the correlation of the size of whole proteome and extracytoplasmic proteins, were undertaken based on different strategies used to identify the bacterial secretome *in silico*.

Table 4.9: The proteome fractions of *Bacillus*' putative extracytoplasmic proteins. The predicted extracytoplasmic proteins of 24 *Bacillus* subspecies/strains were classified into 5 categories according to the Gram-positive classification workflow (see Figure 4.3). Asterisks indicate pathogenic strains. ExCyt= Extracytoplasmic proteins, Lipoprotein= putative lipoprotein, multiple-TM= proteins carrying at least 2 alpha-helical transmembrane regions, one-TM= proteins with one alpha-helix region, Sec-pathway=secreted proteins (carrying Sec signal peptides and not TM proteins), ExtDom protein= proteins carrying extracellular domain(s) and not yet classified in any category.

Name	Proteome size	Total predicted Excyt (%)	multiple-TM	one-TM	Lipoprotein	Sec-pathway	ExtDom protein
Cereus group							
B. anthracis str. A0248	5291	1842 (34.8)	1098	204	143	365	29
B. anthracis str. Ames	5311	1870 (35.2)	1126	207	140	369	30
B. anthracis str. 'Ames Ancestor'	5584	1963 (35.2)	1166	225	146	398	31
B. anthracis str. CDC 684	5902	2060 (34.9)	1204	247	153	416	40
B. anthracis str. Sterne	5287	1889 (35.7)	1182	176	143	347	36
B. cereus 03BB102	5621	2002 (35.6)	1172	235	155	403	36
B. cereus AH187	5758	1998 (34.7)	1170	242	156	405	40
B. cereus AH820	5810	2043 (35.2)	1207	234	119	457	32
B. cereus ATCC 10987	5844	2059 (35.2)	1181	289	149	410	36
B. cereus ATCC 14579	5255	1823 (34.7)	1092	196	145	345	44
B. cereus B4264	5408	1923 (35.6)	1126	207	163	400	35
B. cereus E33L	5641	2049 (36.3)	1243	189	162	419	36
B. cereus G9842	5857	2001 (34.2)	1142	245	166	421	39
B. cereus Q1	5488	1888 (34.4)	1141	191	135	385	42
B. cytotoxicus NVH 391-98	3844	1275 (33.2)	757	116	96	288	24
B. thuringiensis serovar konkukian str. 97-27	5197	1913 (36.8)	1192	163	158	367	36
B. thuringiensis str. Al Hakam	4798	1652 (34.4)	971	166	133	351	37
B. weihenstephanensis KBAB4	5653	2005 (35.5)	1191	201	173	409	33
Non-cereus group							
B. amyloliquefaciens FZB42	3693	1206 (32.7)	734	102	78	256	19
B. clausii KSM-K16	4096	1343 (32.8)	838	99	143	232	23
B. halodurans C-125	4066	1315 (32.3)	767	160	106	235	31
B. licheniformis ATCC 14580 (DSM 13)	4196	1390 (33.1)	852	119	95	287	28
B. licheniformis ATCC 14580 (DSM 13) (2)	4178	1402 (33.6)	840	119	98	316	28
B. pumilus SAFR-032	3681	1234 (33.5)	729	133	86	264	21
B. subtilis subsp. subtilis str. 168	4105	1381 (33.6)	829	111	94	306	27

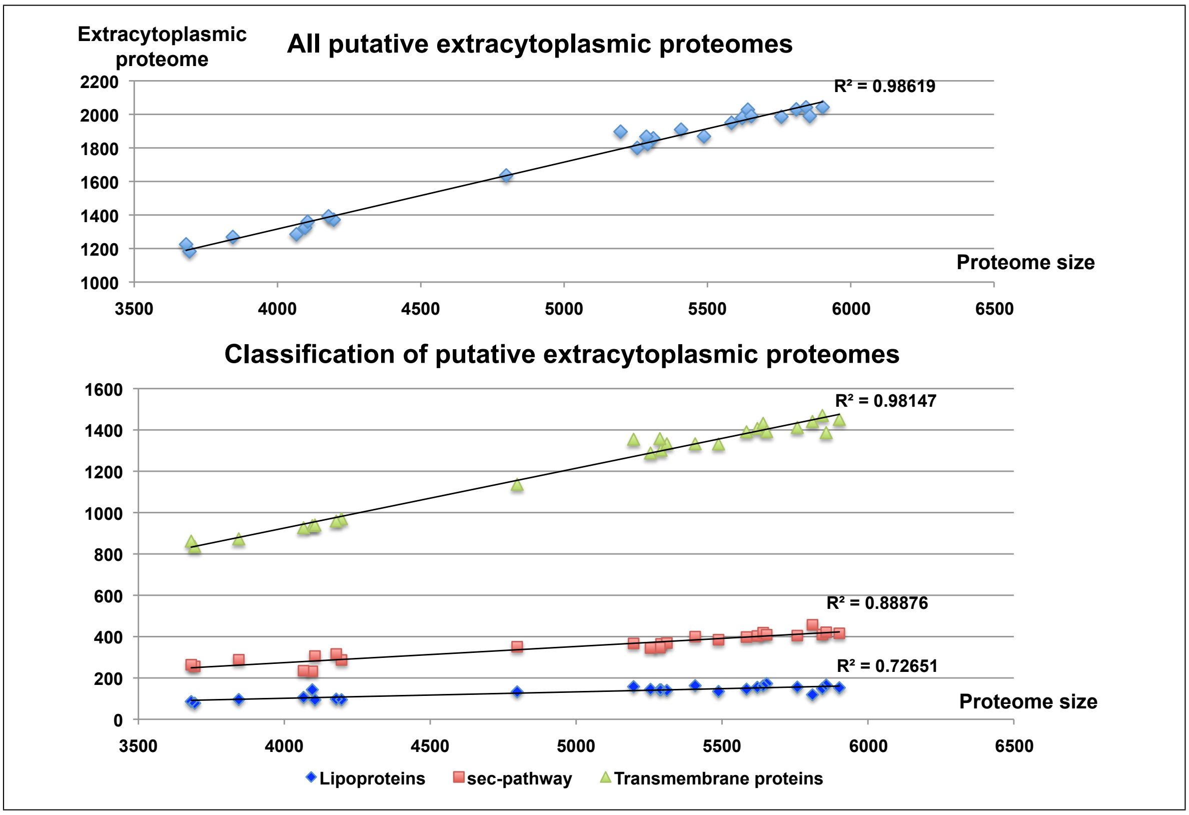


Figure 4.10: The proportions of extracytoplasmic protein classes among all protein sequences across all *Bacillus* species included in the analysis. Linear correlations (R^2) were calculated using Pearson's correlation coefficient.

It appeared that there was a statistically significant difference in the extracytoplasmic proteome size between the cereus group and the non-cereus group (p-value < 0.05) which was probably because the members of cereus generally have a larger proteome size. Unquestionably, the extracytoplasmic proteome size of the pathogenic *Bacillus* was considerably greater in size than the non-pathogenic *Bacillus* species with a statistically significant (p-value < 0.05) (see Figure Figure 4.11). This finding corresponds with a previous experimental study indicating the abundance of surface proteins in *Bacillus* strains with S-layers [Mignot *et al.*, 2001]. Again, our finding disputed the study by Song et al [Song *et al.*, 2009], finding that there was no correlation between secretome size of Gram-positive bacteria and pathogenicity.

It is also important to note that 'pathogenicity' or 'virulence' are terms used to describe microorganisms that are known to be able to disturb normal host physiology, or cause malfunctions in the host body. There is still considerable ambiguity over the meaning of the term [Casadevall and Pirofski, 2001]. Moreover, some microorganisms that are generally considered as non-pathogenic strains may actually cause disease to other organisms and these links may have not yet been discovered [Holden *et al.*, 2004]. In this study, the term 'pathogenicity' was assigned to *Bacillus spp.* based on the genome informa-

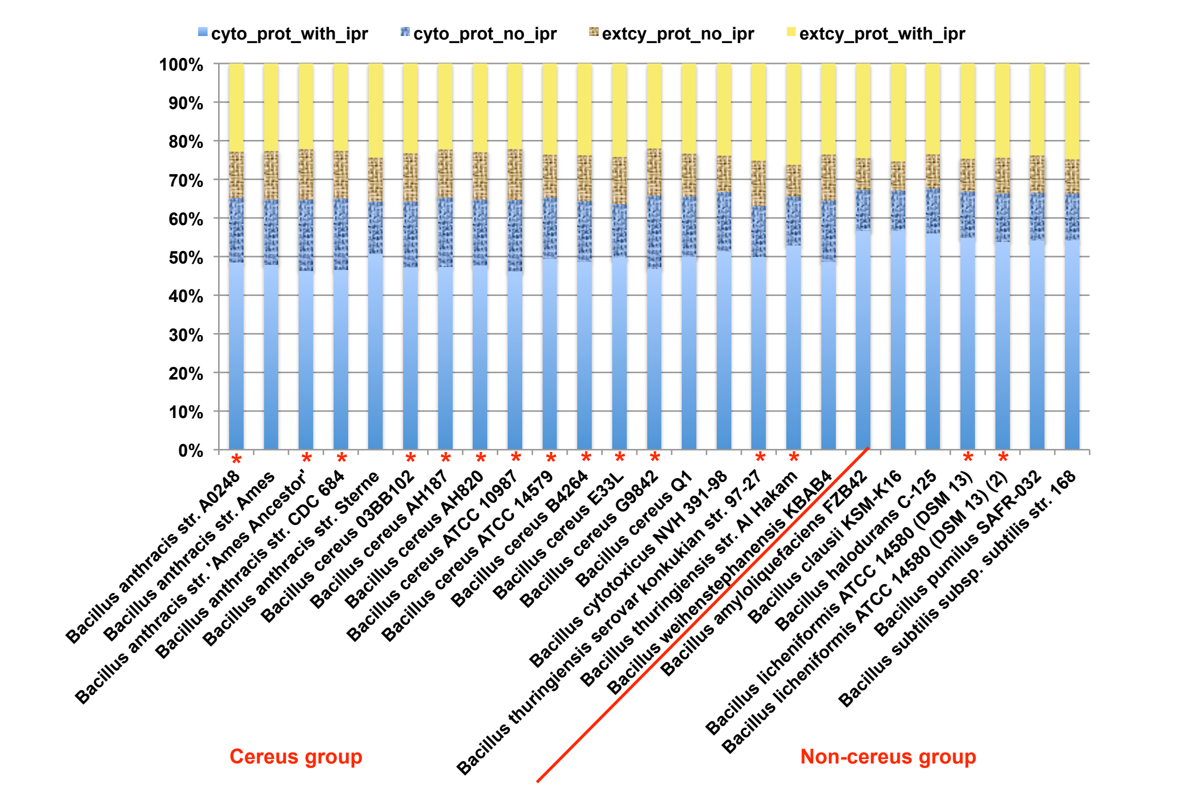


Figure 4.11: The proportion of predicted cellular location of proteins across *Bacillus* species. Each bar represents the entire proteome of a strain. The proportion of predicted extracytoplasmic proteins versus predicted cytoplasmic proteins are shown in yellow and blue shading, respectively. The dotted regions of each bars indicate the proportion of proteins without annotated InterPro domains. Asterisks indicate known pathogenic strains.

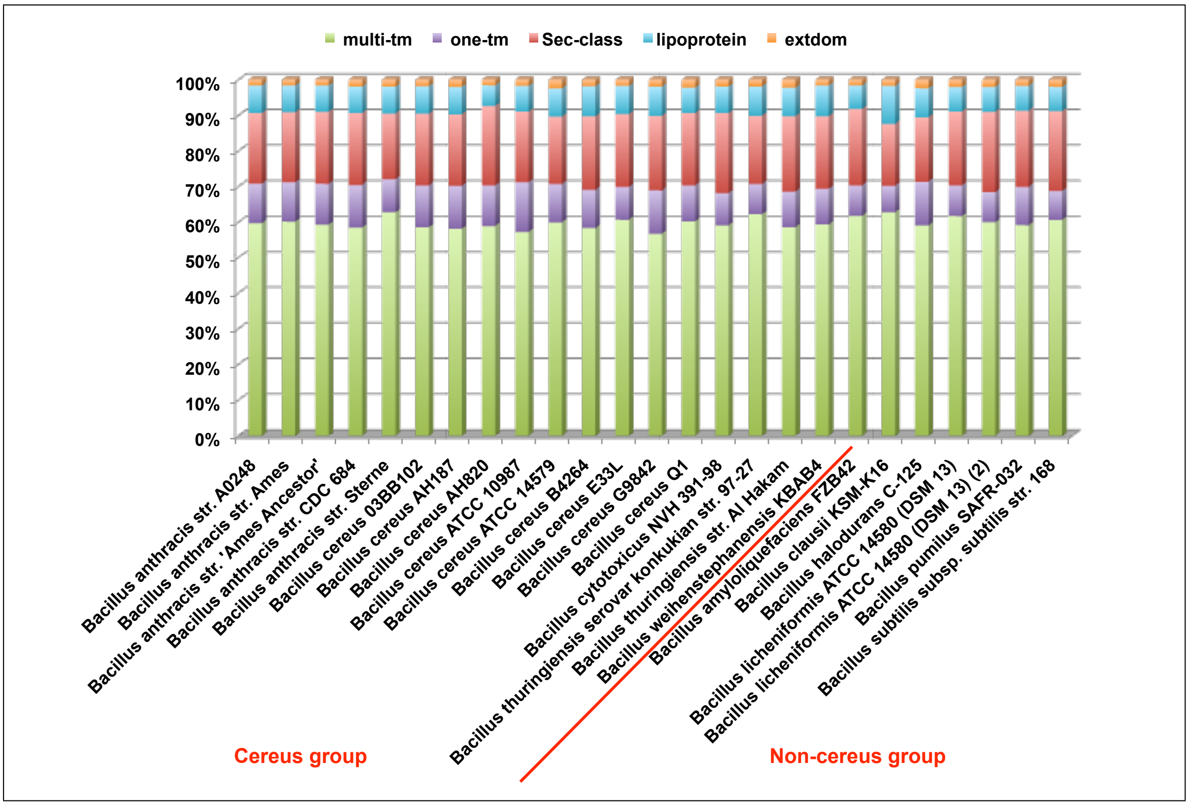


Figure 4.12: The proportion of extracytoplasmic classes of predicted extracytoplasmic proteins across all *Bacillus* species. Each bar represents the entire predicted extracytoplasmic proteome of a strain. The proportion of extracytoplasmic classes of proteins resulted from the Gram-positive protein classification workflow are shown.

tion derived from the GOLD database[2] by considering whether a strain is capable of causing diseases (Table 4.8).

Bacillus extracytoplasmic proteome

In this section, the differences across the Bacilli extracytoplasmic proteome data sets was investigated in more detail, in particular, the contrast between the *Bacillus cereus* group and the non-cereus group. The majority of the *Bacillus* extracytoplasmic proteomes were predicted to have a helix transmembrane region and N-terminal signal peptide cleavage sites by either SPI or SPII. The rest appeared to possess well-known Gram-positive surface-associated motifs or domains including: S-layer motif (PF00395), LPXTG anchoring motif (PF00746, PS50847), LysM domain (PF01476), or other known-characterised extracellular protein signatures such as LRR motif (PF00560), NLPC P60 (PF00877), putative cell wall binding repeat (PF01473). A summary of number of these domains found across *Bacillus* strains are shown in Table 4.10. Interestingly, some of these surface-associated domains are significantly over-represented across the members of the *Bacillus cereus* group, suggesting a larger surface proteome among the cereus group than the non-cereus group. The S-layer homology domain (SLH; PF00395 or InterPro (IPR)001119) not only co-occurs significantly with the members of cereus group (p-value 1 $\times 10^{-4}$), it also appears to be enhanced among the cereus species (p-value 1.76 $\times 10^{-29}$). Likewise, the surface protein from proteins with the Gram-positive bacterial LPXTG anchoring domain (PS50847 or IPR001899) are enriched within the cereus group (p-value 2.73 $\times 10^{-14}$).

Bacillus S-layer homology domain protein

The Surface Layer Homology (SLH) domains mediate association of SLH-domain-bearing proteins non-covalently to the polymers of the secondary cell wall of Gram-positive bacteria [Lee *et al.*, 2003] [Schäffer and Messner, 2005]. These typical essential cell wall polymers such as teichoic, teichuronic acids, lipoteichoic acids or lipoglycans, serve as an anchoring structure for the SLH motif. Several S-layer proteins have been characterised as virulence factors required for pathogenesis, for instance, internalin from *Listeria monocytogenes* and *B. cereus* and PspA from *Streptococcus pneumoniae* [Navarre and Schneewind, 1999] [Fedhila *et al.*, 2006].

SLH domains were presented across almost all *Bacillus spp.* at a higher proportion compared to other anchoring domains annotated on the proteomes. Interestingly, members of the cereus group

[2] http://www.genomesonline.org, accessed 20th August 2010

	*B. anthracis str. ‘Ames Ancestor’	*B. anthracis str. A0248	*B. anthracis str. CDC 684	*B. cereus 03BB102	*B. cereus AH187	*B. cereus AH820	*B. cereus ATCC 10987	*B. cereus ATCC 14579	*B. cereus B4264	*B. cereus E33L	*B. cereus G9842	B. licheniformis ATCC 14580 (DSM 13)	*B. thuringiensis serovar konkukian str. 97-27	*B. thuringiensis str. Al Hakam	B. amyloliquefaciens FZB42	*B. anthracis str. Ames	*B. anthracis str. Sterne	*B. cereus Q1	B. clausii KSM-K16	*B. cytotoxicus NVH 391-98	B. halodurans C-125	B. pumilus SAFR-032	B. subtilis subsp. subtilis str. 168	*B. weihenstephanensis KBAB4
Thermonuclease domain profile (PS50830)	1	1	1	1	1	1	1	0	0	0	2	2	0	0	0	0	0	1	0	0	0	1	3	1
M protein motif (PF00746)	6	6	8	9	13	9	10	12	11	11	11	2	9	10	1	6	8	11	2	5	6	0	1	9
LPXTG, sortase motif (PF04203, PS50847)	5	5	7	9	15	8	10	11	11	10	13	10	7	10	1	5	7	13	7	5	9	7	3	10
LysM domain (PF01476)	6	6	6	6	6	6	6	5	7	7	6	30	6	6	13	6	6	6	7	6	9	11	16	7
Putative cell wall binding repeat (PF01473)	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
S-layer motif (PF00395)	24	20	24	23	23	23	17	10	10	22	11	0	19	21	0	21	21	22	0	19	0	1	0	13
LRR motif (PF00560)	2	1	2	2	2	2	2	2	2	2	2	0	2	2	0	2	2	2	0	0	0	0	0	2
NLPC P60 (PF00877)	4	5	6	5	6	5	5	4	5	5	6	10	5	4	6	3	4	5	3	4	2	5	7	4

Table 4.10: Presence of potential cell surface-associated domains/motifs across *Bacillus* genomes. Asterisks denote the strains from the cereus group.

contain a noticeably greater number of SLH domains (10-24) than the non-cereus group (0-1) (p-value 1.8 $\times 10^{-29}$). These values suggest a phylogenic origin for the presence of S-layer proteins and possibly an ecological pressure. For the *B. anthracis*' surface proteome study, it was suggested that S-layer proteins represent 15% of the cell surface proteins and therefore synthesising S-layer proteins is energy consuming [Fouet, 2009].

The distribution of the SLH domain across the *Bacillus* species immediately raises a question of whether SLH-containing protein sequences carry important protein features evolved among the cereus members or if these surface proteins are involved in host-*Bacillus* interactions, particularly virulence. Therefore, *Bacillus* proteins possessing this SLH cell-surface anchoring domain were investigated in more detail in terms of their molecular functions according to other known features possessed by these proteins. The 338 proteins were found to have at least one SLH domain and at most three domains.

The *Bacillus* proteins with SLH domains are annotated as: putative penicillin-binding domain; cell-wall hydrolase/autolysin; peptidoglycan endo-beta N-acetylglucosaminidase and N-acetylmuramoyl-L-alanine amidase fusion; S-layer protein EA1; S-layer protein Sap precursor; Ig domain-containg protein; crystal protein; internalin; N-acetylmuramoyl-L-alanine amidase; iron transport-associated domain-containing protein; GW repeat-containing protein; NEAr transporter and hypothetical proteins. Internalin from *B. cereus* ATCC14579, was considered as a virulence factor during an infection of the bacteria in insect larvae, *Galleria mellonella* [Fedhila *et al.*, 2006]. The NEAr transporter or NEAT domain, exclusive to Gram-positive bacteria, has a role in heam binding for the acquisition of iron from the host body. The domain was believed to involved in an iron transporter because the NEAT-domain encoding genes were located adjacent to genes coding components of the Fe^{3+} siderophore transporter [Grigg *et al.*, 2007][Andrade *et al.*, 2002]. Some of these *Bacillus* S-layer protein are hypothetical. However, these hypothetical proteins may contain as yet unidentified conserved functional regions that might be important to their survival in specific environments. These proteins are therefore still of interest in terms of how they are an advantage to a particular group of organisms. Particularly, in this case, how these hypothetical proteins could assist in the adaptation of the cereus species to a host body. A novel M60-like protein domain is described later (Chapter 7) that has been found in several hypothetical proteins and is potentially important for several mucosa-associated microorganisms for interacting with their hosts.

Protein domains possessed by the S-layer proteins of the cereus group's members is shown in Table 4.11. These functional domains play roles in drug resistance (e.g. Beta-lactamase-related), protein-protein interaction (e.g. Leucine-rich repeat), peptidoglycan catabolic process (e.g N-acetylmuramoyl-

L-alanine amidase, family 2), membrane transport (e.g. NEAr transporter), and pathogenesis (e.g. Immunoglobulin E-set).

Several protein domains found on the S-layer proteins also appear to be over-represented among the cereus species (marked by '*' in Table 4.12).

Domain description	Bant AA	Bant A0248	Bant Ames	Bant CDC 684	Bant Sterne	Bcer 03BB102	Bcer AH187	Bcer AH820	Bcer ATCC 10987	Bcer ATCC 14579	Bcer B4264	Bcer E33L	Bcer G9842	Bcer Q1	Bcyt NVH 391-98	Bwei	Bthu konkukian str. 97-27	Bthu Al Hakam	**Grand Total**
Hydrolase activity																			
Beta-lactamase-like	1	1	1	1	1	1	1	1	-	-	-	1	-	1	-	-	-	1	11
Beta-lactamase-related	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	1	2
Beta-lactamase-type transpeptidase fold	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	1	2
Cell wall hydrolase/autolysin, catalytic	4	4	3	4	3	3	2	3	2	2	1	3	2	2	3	2	2	3	48
Glycoside hydrolase, catalytic core	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	1
Lysozyme subfamily 2	1	1	1	1	1	1	-	1	-	-	2	-	2	1	-	-	-	1	13
Mannosyl-glycoprotein endo-beta-N-acetylglucosamidase	1	1	1	1	1	1	1	1	1	2	2	1	2	1	2	1	1	1	22
N-acetylmuramoyl-L-alanine amidase, family 2	3	1	3	3	3	3	3	3	3	2	3	2	2	3	2	3	2	3	47
Adhesion/Cell binding																			
Cell wall/choline-binding repeat	-	-	-	-	-	-	1	1	-	-	-	1	-	-	-	-	-	-	3
Putative cell wall binding repeat	-	-	-	-	-	-	1	1	-	-	-	1	-	-	-	-	-	-	3
Fibronectin, type III	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	1
Leucine-rich repeat	1	-	1	1	1	1	1	1	1	1	1	1	1	1	-	1	1	1	16
Leucine-rich repeat, typical subtype	-	-	-	1	-	1	1	1	-	-	1	-	1	1	-	-	-	-	7
Protein folding																			
Chaperonin Cpn60, conserved site	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	1	-	-	2
Prefoldin	1	1	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	4
DNA/RNA metabolism																			
Guanine-specific ribonuclease N1 and T1	-	-	-	-	-	-	1	1	-	-	-	-	-	-	-	-	-	-	2
Ribonuclease/ribotoxin	-	-	-	-	-	-	1	1	-	-	-	-	-	-	-	-	-	-	2
Transport activity																			
NEAr transporter	1	1	1	1	1	1	2	2	2	1	1	2	1	2		2	2	2	25
Lipocalin	1	1	1	1	1	1	1	1	1	-	-	1	-	1	-	1	-	1	13
Others																			
Peptidase, cysteine peptidase active site	-	-	-	-	-	-	1	-	1	-	-	-	-	1	-	-	-	-	3
SH3, type 3	1	-	1	1	1	1	2	2	1	1	2	2	1	1	-	1	2	1	21
Bacterial SH3-like region (mediate many diverse processes such as increasing local concentration of proteins, altering their subcellular location and mediating the assembly of large multiprotein complexes)	1	-	1	1	1	1	1	1	-	1	1	1	-	1	-	1	2	1	15
Excalibur calcium-binding domain	1	1	1	1	1	1	1	1	1	-	-	1	-	1	1	1	1	1	15
Bacterial Ig-like, group 2	1	1	1	1	1	1	1	-	-	-	-	1	-	-	1	-	-	-	9
Immunoglobulin E-set	1	-	1	1	1	1	1	1	1	1	1	1	1	1	-	1	1	1	16
Transglutaminase-like	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	18
YkuD domain	-	-	-	-	-	1	-	-	-	-	-	1	-	-	-	-	-	-	2

Domain description	Bant AA	Bant A0248	Bant Ames	Bant CDC 684	Bant Sterne	Bcer 03BB102	Bcer AH187	Bcer AH820	Bcer ATCC 10987	Bcer ATCC 14579	Bcer B4264	Bcer E33L	Bcer G9842	Bcer Q1	Bcyt NVH 391-98	Bwei	Bthu konkukian str. 97-27	Bthu Al Hakam	**Grand Total**
Protein of unknown function DUF187	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	1

Table 4.11: Protein domain composition of S-layer proteins of the Baciilus species in the cereus group. The number in each cell represents the number of particular protein domains found in a specific *Bacillus* strain. Some of the domains are known to involved in hydrolase activity, adhesion, protein folding, DNA and RNA metabolisms, as well as transport activity.

4.3.4 Identification of extracytoplasmic protein domains shared or predominant among *Bacillus* species

This section demonstrates the use of data obtained from the project's Microbase workflows (described in Chapter 3) and the workflows described earlier in this chapter. The aim of the work presented here was to reveal important protein features that are potentially involved in host-microbe interactions and the adaptation to the respective life-styles of the different *Bacillus spp*. The focus was placed particularly on features of the *Bacillus*' extracytoplasmic proteome because they are microbial components that interface with the host environment.

Approximately, 20-25% of the identified protein-coding gene sequences did not contain any known InterPro protein domains published at the time of the study (see Figure 4.11). Within the proportion of proteins with no identified conserved regions, roughly 8-13% are predicted *Bacillus* extracytolasmic proteins. To examine whether the current protein domain profiles of known conserved protein features can be used to distinguish different *Bacillus* species included in this study, the relationship among *Bacillus* species based on their protein domain profiles was investigated. If the domain profile-based relationship reflected the known model of phylogenetic relationships among *Bacillus* species based on 16S rDNA, it could be concluded that different phenotypes of *Bacillus* must be determined by the presence of some of these annotated protein domains.

As a result, a hierarchy clustering of all *Bacillus* species using a number of InterPro domains annotated by InterProScan on each *Bacillus* proteome data set was generated (see Figure 4.13). Interestingly, the relationship of *Bacillus* with respect to the domain profiles corresponds to the phylogenetic tree generated using 16S rDNA sequences (see Figure 4.9). Like the 16S rDNA phylogenetic tree, the members of the *Bacillus cereus* group were clustered in one clade and separated clearly from the non-cereus species. This result indicates that, for the *Bacillus* species, InterPro domain composition can be used as a guideline for their phylogenetic relationships. This is not surprising as protein sequences and their functional components are products of gene-coding DNA sequences. However, a sufficient number of known protein signatures among the proteome data set is required for a reliable suggestion of the phylogenetic relationship. In this case, it seemed that the *Bacillus*' InterPro domain profiles have a satisfactory level of information that can be employed for further investigation of the differences in the *Bacillus* phenotypes or their ability to thrive in different environments.

In the next step, the analysis to identify protein domains (InterPro entries) that discriminate the bacteria of the cereus group from the non-cereus group was performed. The hypergeometric test (see Section 2.9.1) was used as a significance test to evaluate the probability that an InterPro domain

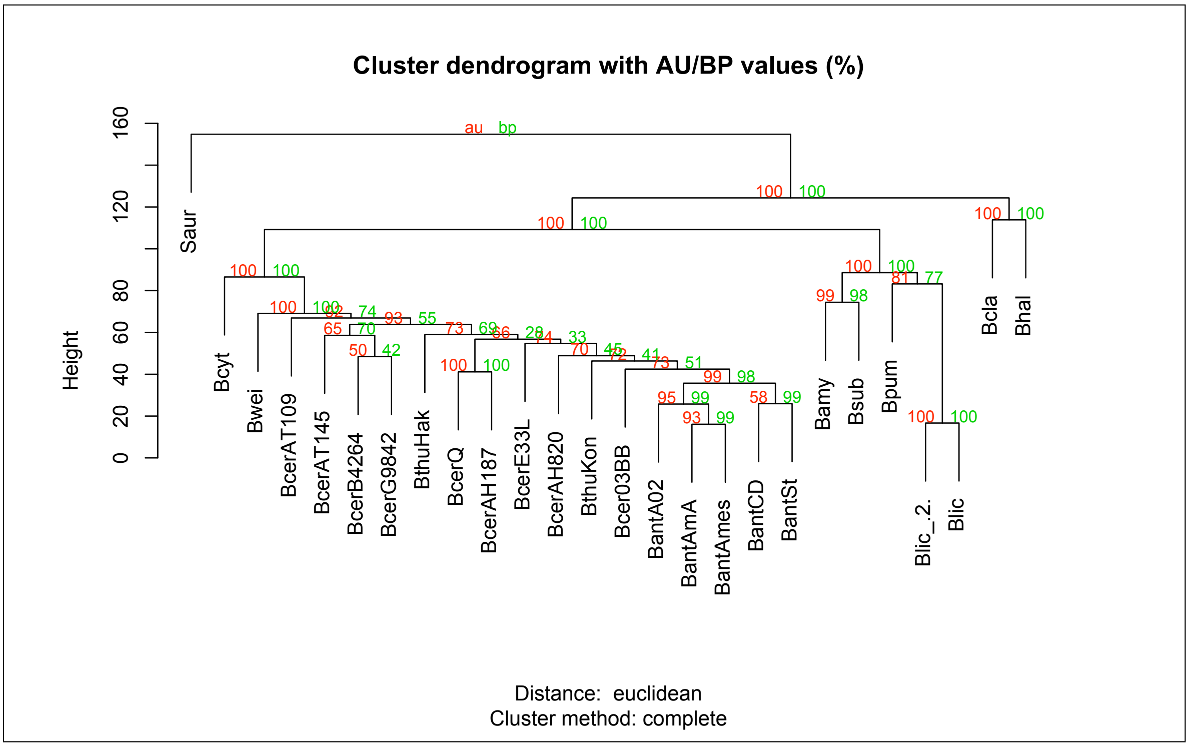


Figure 4.13: Relationship of *Bacillus spp*. based on their normalised Interpro domain composition. *Staphylococcus aureus*'s data set was used as an outgroup. The hierarchical cluster was computed using the complete linkage clustering method and Euclidean distance based similarity. Values on the edges of the clustering are p-values (%). Red values are AU (Approximately Unbiased) p-values, and green values are BP (Bootstrap) values. BantAmes= *B. anthracis* str. Ames, Bamy= *B. amyloliquefaciens* FZB42, BantA02= *B. anthracis* str. A0248, BantAmA= *B. anthracis* str. 'Ames Ancestor', BantCD= *B. anthracis* str. CDC684, BantSt= *B. anthracis* str. Sterne, Bcer03BB= *B. cereus* 03BB102, BcerAH187= *B. cereus* AH187, BcerAH820= *B. cereus* AH820, BcerAT109= *B. cereus* ATCC10987, BcerAT145= *B. cereus* ATCC14579, BcerB4264= *B. cereus* B4264, BcerE33L= *B. cereus* E33L, BcerG9842= *B. cereus* G9842, BcerQ= *B. cereus* Q1, Bcla= *B. clausii* KSM-K16, Bcyt= *B. cytotoxicus* NVH391-98, Bhal= *B. halodurans* C-125, Blic= *B. licheniformis* ATCC14580 (DSM13), Bpum= *B. pumilus* SAFR-032, Bsub= *B. subtilis* subsp. subtilis str. 168, BthuHak= *B. thuringiensis* str. Al Hakam, BthuKon= *B. thuringiensis* serovar konkukian str. 97-27, Bwei= *B. weihenstephanensis* KBAB4

occur preferentially among members of one group. The hypergeometric mean value was calculated to determine the direction of these associations (i.e associated to cereus or non-cereus group). This statistical technique was applied to identify the association of both the co-occurrence and the abundance of a domain in the *Bacillus* groups. The co-occurrence evaluation considers the presence and absences of a domain in the bacterial groups, whereas the abundance (enrichment) analysis takes into account the number of domains among members of each *Bacillus* group. Pearson's correlation coefficient (see Section 2.9.2) was also employed to measure a linear relationship between a domain and the two *Bacillus* groups. These statistical techniques were applied to all 3,078 InterPro entries annotated on at least one of a *Bacillus* protein-coding gene sequence. The entries found to be over-represented among the cereus group as well as expressed on the extracytoplasmic proteins were shown (see Table 4.12). The distribution of these domains across *Bacillus* species is illustrated in a heatmap (see Figure 4.14).

Not surprisingly, domains known to be involved in pathogenesis or virulence were found significantly among the cereus group members, for instance, Bacillus PapR (IPR09239), hemolysin. PapR peptides promote the expression of the PlcR regulon, a regulator that activates various virulence factors in *B. cereus* and *B. thuringiensis* [Slamti and Lereclus, 2002]. Other known virulence-associated domains predominant across cereus members included Leukocidin / porin (IPR016183), Leukocidin / haemolysin (IPR01340), and thiol-activated cytolysin (IPR001869). Many other domains function as proteases and some of these are known as virulence factors e.g. thermolysin (IPR013856, IPR001570), viral enhancin, collagenase (IPR013510, IPR013661), fungalysin (IPR001842), archaeal and bacterial peptidase (IPR007280), peptidase M4 and M36 (IPR011096), and peptidase S15 (IPR000383, IPR013736). Several antibiotic-resistance protein domains were also identified including penicillin amidase (IPR002692), and Beta-lactamase class B (IPR001018). As discussed earlier (Section 4.3.3), several bacterial cell wall anchor were identified including SLH (discussed previously, Section 4.3.3), and surface protein from Gram-positive cocci (IPR001899).

In addition, domains known to be involved in bacterial adhesion to mammal or insect or plant surfaces were also predominant among the cereus bacteria. These domains include bacterial adhesion (IPR008966), collagen-binding surface protein Cna-like (IPR008454), bacterial cellulose-binding family II (IPR001919), bacterial pullanase-associated protein (IPR005323), and chitin-binding domain 3 (IPR004302). Domains regarded as being involved in transport were also found, such as amino acid transporter (IPR013057), short chain fatty acid transporter (IPR006160), TGF-beta receptor, type I/II (IPR018456, IPR000109), Branched-chain amino acid transport system II carrier protein (IPR004685), Ferrous iron transport protein (IPR011619, IPR011640), ATPase K+ transporter (IPR004623, IPR003820), Zinc/iron permease (IPR003689), Nicotinamide mononucleotide transporter (IPR006419), and Ion transport 2 (IPR013099). Some domains serve as a binding site for protein-protein interaction (e.g. WD40 repeat (IPR001680) and Leucine-rich repeat (IPR001611)). Using our approach, we have identified several proteins of unknown function that are dominantly presented among the cereus group (see Table 4.13). These unknown-function domains could be subject to further study for their specific involvement in host-microbe interactions or particular environment adaptations of the cereus bacteria.

Table 4.12 **List of InterPro (IPR) entries that are overrepresented across extracytoplasmic proteins of the *Bacillus cereus* group's members compared to those in the non-cereus group.** The IPR entries listed here are domains with known functions and annotated mainly on putative extracellular proteins of the *Bacillus cereus* group (> 50%). 'extprot' shows fractions of putative extracytoplasmic proteins that possess a particular domain in all proteins annotated with the domain in the cereus's species. An asterisk denotes domains found on S-layer proteins. These domains have either a significant co-occurrence p-value ('co_pvalue') or an abundance p-value ('pvalue') of ≤ 0.01, therefore the null hypothesis of no association is rejected. The p-values (uncorrected) were calculated using the hypergeometric test. 'pvalue' indicates the probability of the abundance of a domain within the cereus group, whereas 'co_pvalue' denotes the probability that a domain is present in members of the cereus group. 'corr' represents correlation scores measuring a linear relationship between numbers of a given domain and the two groups of *Bacillus spp*. The closer the score to 1, the higher the strength of the linear correlation between that domain and the cereus bacteria. The correlation score was computed using Pearson's correlation coefficient. This score reflects the over-representation of a domain among the cereus group. The distribution of each domain across *Bacillus* species is shown in Figure 4.14.

InterPro entry	**extprot** (%)	**C**	**NC**	**pvalue**	**co_pvalue**	**corr.**	**description**
Hydrolase activity							
Peptidase S45, penicillin amidase	94	17	0	7.65×10^{-3}	1.66×10^{-5}	0.91	IPR002692
Beta-lactamase, class B, conserved site	88	17	1	3.43×10^{-2}	2.62×10^{-4}	0.80	IPR001018
Beta-lactamase-related	61	291	29	2.47×10^{-13}	1.00	0.89	IPR001466*
Phospholipase C/P1 nuclease, core	100	18	0	5.74×10^{-3}	2.08×10^{-6}	1.00	IPR008947
Phospholipase C, zinc-binding, prokaryotic	100	18	0	5.74×10^{-3}	2.08×10^{-6}	1.00	IPR001531
Phospholipase C, phosphatidylinositol-specific , X region	100	19	2	5.62×10^{-2}	2.13×10^{-3}	0.63	IPR000909
N-acetylmuramoyl-L-alanine amidase, family 2	51	109	15	1.90×10^{-4}	2.80×10^{-1}	0.79	IPR002502*
Transport activity							
Protein transport							
Amino acid transporter, transmembrane	100	21	0	2.43×10^{-3}	1.65×10^{-3}	0.57	IPR013057
Branched-chain amino acid transport system II carrier protein	100	114	9	3.11×10^{-7}	2.80×10^{-1}	0.95	IPR004685
Ferrous iron transport							
Ferrous iron transport protein B, C-terminal	100	32	2	3.61×10^{-3}	3.95×10^{-4}	0.84	IPR011640
Ferrous iron transport protein B, N-terminal	100	47	2	1.03×10^{-4}	3.95×10^{-4}	0.84	IPR011619
NEAr transporter	96	91	4	5.73×10^{-8}	2.13×10^{-3}	0.83	IPR006635*
Potassium transporting ATPase							
ATPase, K+ transporting, A subunit	100	18	0	5.74×10^{-3}	2.08×10^{-6}	1.00	IPR004623
ATPase, K+ transporting , KdpC subunit	100	18	0	5.74×10^{-3}	2.08×10^{-6}	1.00	IPR003820
Zinc/iron permease	100	18	0	5.74×10^{-3}	2.08×10^{-6}	1.00	IPR003689
Other transports							
Short chain fatty acid transporter	100	17	2	8.12×10^{-2}	2.13×10^{-3}	0.69	IPR006160
AmiS/UreI transporter	100	18	0	5.74×10^{-3}	1.66×10^{-5}	0.85	IPR003211
Anaerobic c4-dicarboxylate membrane transporter	100	18	0	5.74×10^{-3}	2.08×10^{-6}	1.00	IPR004668
Nicotinamide mononucleotide transporter PnuC	100	14	0	1.81×10^{-2}	6.86×10^{-4}	0.70	IPR006419
PhoU	50	36	5	2.38×10^{-2}	2.77×10^{-3}	0.78	IPR008170
Killing of cells of another organism							
Leukocidin/porin	100	18	0	5.74×10^{-3}	3.57×10^{-3}	0.54	IPR016183
Leukocidin/haemolysin	100	16	0	1.02×10^{-2}	3.57×10^{-3}	0.54	IPR001340
Peptidase activity							
Peptidase M36, fungalysin	100	18	0	5.74×10^{-3}	2.08×10^{-6}	1.00	IPR001842
Peptidase M9A/M9B, collagenase C-terminal	100	48	0	1.06×10^{-6}	2.08×10^{-6}	0.80	IPR013510
Peptidase M9A/M9B, N-terminal	100	46	0	1.87×10^{-6}	2.08×10^{-6}	0.87	IPR013661
Peptidase S15	88	17	0	7.65×10^{-3}	1.66×10^{-5}	0.91	IPR000383
Propeptide, peptidase M4 and M36	96	108	3	1.22×10^{-10}	3.95×10^{-4}	0.93	IPR011096
Peptidase M4, thermolysin C-terminal	100	90	3	1.25×10^{-8}	3.95×10^{-4}	0.92	IPR001570
Peptidase M4, thermolysin	100	89	3	1.61×10^{-8}	3.95×10^{-4}	0.91	IPR013856

InterPro entry	**extprot** (%)	**C**	**NC**	**pvalue**	**co_pvalue**	**corr.**	**description**
Peptidase, archaeal and bacterial C-terminal	100	61	3	1.64×10^{-5}	2.77×10^{-3}	0.82	IPR007280
Peptidase S15/CocE/NonD, C-terminal	82	17	0	7.65×10^{-3}	1.66×10^{-5}	0.91	IPR013736
Peptidase M60, viral enhancin protein	92	12	0	3.21×10^{-2}	7.14×10^{-3}	0.52	IPR004954
Binding							
Carbohydrate binding							
Cellulose-binding, family II, bacterial type	100	18	0	5.74×10^{-3}	2.08×10^{-6}	1.00	IPR001919
Bacterial pullanase-associated protein	100	17	1	3.43×10^{-2}	2.62×10^{-4}	0.80	IPR005323
Chitin-binding, domain 3	100	42	5	8.71×10^{-3}	7.00×10^{-2}	0.80	IPR004302
Protein binding							
Collagen-binding surface protein Cna-like, B region	97	76	6	2.89×10^{-5}	7.00×10^{-2}	0.70	IPR008454
TIMP-like, OB-fold	100	14	0	1.81×10^{-2}	6.86×10^{-4}	0.70	IPR008993
Leucine-rich repeat	100	33	0	7.78×10^{-5}	1.66×10^{-5}	0.89	IPR001611*
Cholesterol binding							
Thiol-activated cytolysin	100	17	0	7.65×10^{-3}	1.66×10^{-5}	0.91	IPR001869
Cell surface binding							
S-layer homology region	100	338	1	6.36×10^{-41}	3.95×10^{-5}	0.90	IPR001119*
Surface protein from Gram-positive cocci, anchor region	100	183	17	1.64×10^{-9}	1	0.84	IPR001899
Cell adhesion							
Adhesion, bacterial	84	86	6	3.35×10^{-6}	7.00×10^{-2}	0.71	IPR008966
Biosynthetic process							
Fatty acid hydroxylase	100	18	0	5.74×10^{-3}	1.66×10^{-5}	0.85	IPR006694
Poly-beta-hydroxybutyrate polymerase, N-terminal	94	18	0	5.74×10^{-3}	2.08×10^{-6}	1.00	IPR010941
Others							
Integral membrane protein 1906	100	18	0	5.74×10^{-3}	2.08×10^{-6}	1.00	IPR010178
Glycerophosphoryl diester phosphodiesterase, membrane domain	100	18	0	5.74×10^{-3}	2.08×10^{-6}	1.00	IPR018476
Flagellar basal body FlaE	100	18	0	5.74×10^{-3}	2.08×10^{-6}	1.00	IPR011491
Bacillus PapR	100	17	0	7.65×10^{-3}	1.66×10^{-5}	0.91	IPR009239
Acid phosphatase (Class B)	100	17	0	7.65×10^{-3}	1.66×10^{-5}	0.91	IPR005519
YhhN-like	100	18	0	5.74×10^{-3}	2.08×10^{-6}	1.00	IPR012506
Lysylphosphatidylglycerol synthetase/UPF0104	100	18	1	2.72×10^{-2}	3.95×10^{-5}	0.90	IPR005242
GPR1/FUN34/yaaH	100	17	1	3.43×10^{-2}	2.62×10^{-4}	0.80	IPR000791
Glutaredoxin active site	100	17	0	7.65×10^{-3}	1.66×10^{-5}	0.91	IPR011767
PKD	100	43	1	4.85×10^{-5}	3.95×10^{-5}	0.86	IPR000601
Respiratory-chain NADH dehydrogenase, subunit 1, conserved site	100	18	0	5.74×10^{-3}	2.08×10^{-6}	1.00	IPR018086
Respiratory-chain NADH dehydrogenase, subunit 1	100	18	0	5.74×10^{-3}	2.08×10^{-6}	1.00	IPR001694
NADH-ubiquinone/plastoquinone oxidoreductase, chain 6	100	18	0	5.74×10^{-3}	2.08×10^{-6}	1.00	IPR001457
Ionotropic glutamate receptor	100	46	0	1.87×10^{-6}	2.50×10^{-4}	0.67	IPR001320
HPP	100	15	0	1.36×10^{-2}	2.50×10^{-4}	0.76	IPR007065
NERD	64	25	2	1.68×10^{-2}	6.68×10^{-3}	0.59	IPR011528
NADH-ubiquinone/plastoquinone oxidoreductase, chain 3	100	17	0	7.65×10^{-3}	1.66×10^{-5}	0.91	IPR000440
PepSY-associated TM helix	100	17	3	1.35×10^{-1}	2.13×10^{-3}	0.47	IPR005625
Transcription factor TFIIB related	100	14	0	1.81×10^{-2}	6.86×10^{-4}	0.70	IPR000812
Membrane bound O-acyl transferase, MBOAT	100	44	5	6.10×10^{-3}	7.00×10^{-2}	0.72	IPR004299
WD40 repeat	54	79	9	3.09×10^{-4}	7.00×10^{-2}	0.80	IPR001680
Bacterial SH3-like region	88	184	18	3.77×10^{-9}	2.80×10^{-1}	0.89	IPR003646*
SH3, type 3	87	188	20	1.22×10^{-8}	2.80×10^{-1}	0.86	IPR013247*
L-lactate dehydrogenase, active site	96	54	7	4.93×10^{-3}	1.00	1.00	IPR018177

Table 4.12

Table 4.13: List of unknown-function InterPro entries predominant across extracytoplasmic proteins of the *Bacillus cereus* group's members compared to those in the non-cereus group. See Table 4.12 for a detailed description.

InterPro entry	extprot (%)	C	NC	pvalue	co_pvalue	corr.	description
IPR010380	100	46	0	1.87×10^{-6}	2.08×10^{-6}	0.85	Protein of unknown function DUF975
IPR010390	100	38	0	1.86×10^{-5}	2.08×10^{-6}	0.93	Protein of unknown function DUF990
IPR010539	100	19	0	4.31×10^{-3}	2.08×10^{-6}	0.93	Protein of unknown function DUF1112
IPR009323	100	18	0	5.74×10^{-3}	2.08×10^{-6}	1.00	Protein of unknown function DUF979
IPR010398	100	18	0	5.74×10^{-3}	2.08×10^{-6}	1.00	Protein of unknown function DUF997
IPR010387	100	18	0	5.74×10^{-3}	2.08×10^{-6}	1.00	Protein of unknown function DUF988
IPR001434	82	73	1	1.5×10^{-8}	3.95×10^{-5}	0.79	Protein of unknown function DUF11
IPR012452	93	69	2	3.95×10^{-7}	3.95×10^{-5}	0.80	Protein of unknown function DUF1657
IPR018639	100	17	0	7.65×10^{-3}	1.66×10^{-5}	0.91	Protein of unknown function DUF2062
IPR011397	100	31	1	1.10×10^{-3}	3.95×10^{-5}	0.86	Uncharacterised conserved protein UCP033101
IPR010374	100	17	0	7.65×10^{-3}	1.66×10^{-5}	0.91	Protein of unknown function DUF969
IPR018383	100	18	1	2.72×10^{-2}	3.95×10^{-5}	0.90	Uncharacterised protein family UPF0324
IPR000612	100	16	0	1.02×10^{-2}	7.49×10^{-5}	0.83	Uncharacterised protein family UPF0057
IPR012963	100	18	4	7.10×10^{-4}	3.95×10^{-4}	0.87	Protein of unknown function DUF1700
IPR007563	100	18	2	6.78×10^{-2}	3.95×10^{-4}	0.80	Protein of unknown function DUF554
IPR007163	100	18	2	6.78×10^{-2}	3.95×10^{-4}	0.80	Protein of unknown function DUF368
IPR006837	100	18	2	6.78×10^{-2}	3.95×10^{-4}	0.80	Protein of unknown function DUF610, YibQ
IPR009825	100	18	1	2.72×10^{-2}	2.62×10^{-4}	0.75	Protein of unknown function DUF1393
IPR003848	76	87	3	2.67×10^{-8}	2.77×10^{-3}	0.98	Protein of unknown function DUF218
IPR009959	100	20	1	1.69×10^{-2}	2.80×10^{-3}	0.60	Protein of unknown function DUF1486
IPR005226	100	17	2	8.12×10^{-2}	2.13×10^{-3}	0.70	Conserved hypothetical protein CHP00245
IPR012873	100	13	0	2.41×10^{-2}	1.65×10^{-3}	0.65	Protein of unknown function DUF1672
IPR009732	100	19	2	5.62×10^{-2}	6.68×10^{-3}	0.56	Protein of unknown function DUF1304

4.4 Discussion

Many algorithms and strategies exist to aid the prediction of protein subcellular localisation. However, every tool has different advantages and disadvantages. To our knowledge, a specific prediction tool is normally not applicable to proteins from all taxa for which genome data exists. The approach used in this study employed several well-known bioinformatics tools to facilitate the prediction of all potential extracytoplasmic protein sequences among the three domains of microbial cellular life. The tools included in the workflow were carefully selected taking the variation of microorganisms' surface structures into consideration. The prediction results from each tool were considered sequentially using appropriate workflows designed to suit each type of microbial cell surface. Constructing a bioinformatics workflow to perform a selective integration of results from various tools has been shown to provide a considerably high performance (positive predictive value 87-100%, sensitivity 86-93%) in the prediction of extracytoplasmic proteins across proteomes from different groups of organisms with distinctive cell-surface structures. Combining results from different targeting signal predictions allows the differentiation of the extracytoplasmic protein sequence according to the presence of their targeting features. For example, in this study, it was possible to distinguish sequences

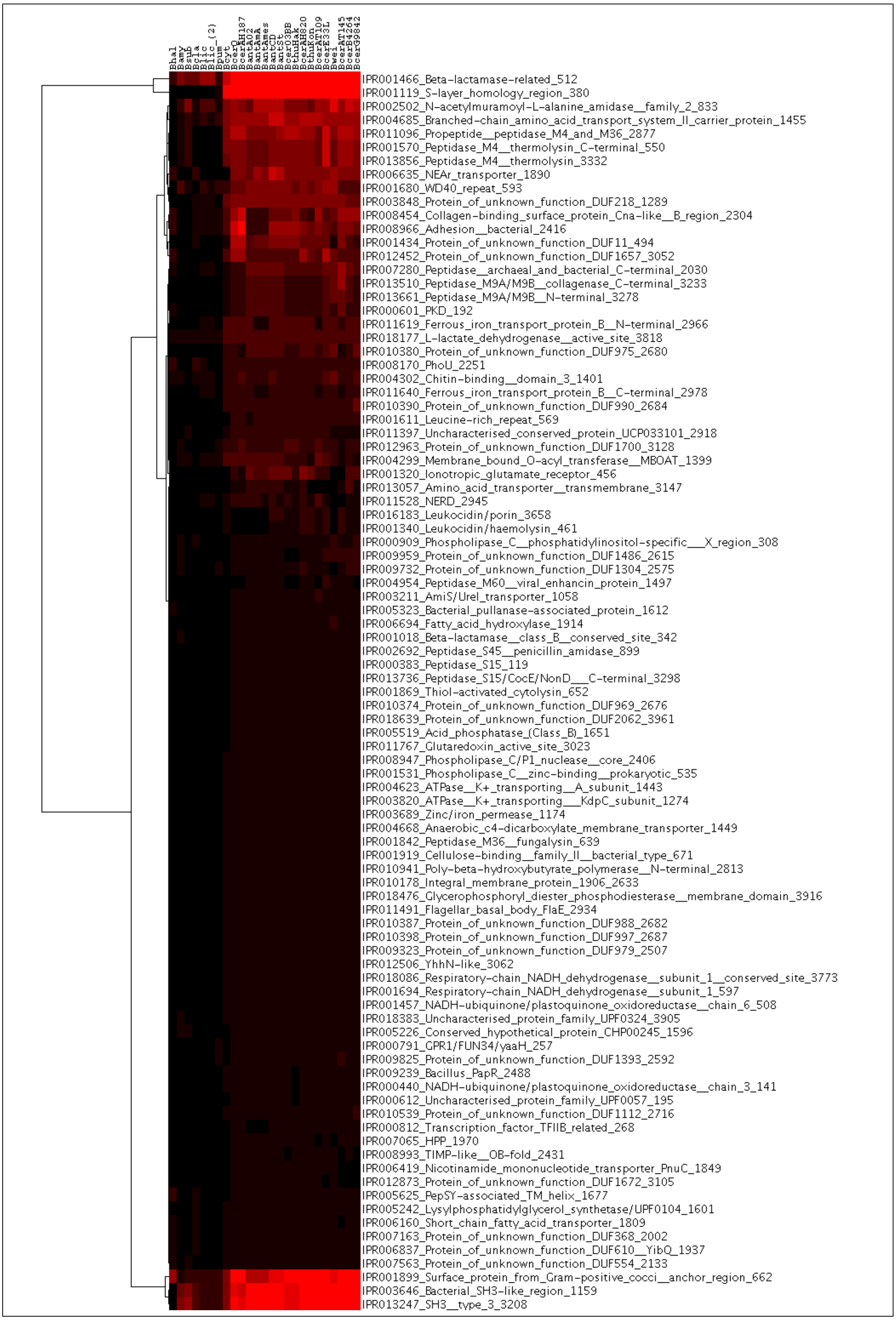


Figure 4.14: A heatmap dendrogram showing the distribution of InterPro protein domains predominant among the members of the *Bacillus cereus* group. Rows represent InterPro domains that are listed in Table 4.12. Columns denote different *Bacillus* species. Each cell is colourised based on the number of protein sequences (unnormalised) possessing that protein domain in that organism's proteome. The brighter the colour, the larger the number.

with alpha-helical transmembrane from lipoproteins. Furthermore, cell-surface anchoring proteins were also identified. This kind of detailed description would not have been feasible with the use of any single prediction tool available at the time of study.

Several limitations were identified in the strategy used for the construction of the identification and classification of extracytoplasmic proteins workflows. For example, the workflow did not utilise any tool specifically designed to predict transmembrane beta-barrels of the Gram-negative bacteria outer membrane proteins (e.g. [Freeman and Wimley, 2010]). Nevertheless, most of the known Gram-negative outer membrane proteins were identified by the presence of the N-terminal signal peptides detectable by SignalP. Moreover, several Gram-negative bacterial outer-membrane proteins possessing known outer membrane domains (listed in Table 3.1) were also classified into a set of putative extracytoplasmic proteins by the workflow.

In addition, it is known that features specific to Mycobacterium surface and secreted proteins might be falsely detected by the standard bacterial protein predictors trained with Gram-positive and Gram–negative bacteria [Rashid *et al.*, 2007]. Therefore, a tool trained specifically with Mycobacterium proteins could be added to the workflow in future. Such a tool may significantly improve the workflow's accuracy.

The workflow did not cover the identification of Gram-negative proteins secreted via the type III secretion system, where sequence patterns are not easily recognisable as there is no clear common sequence pattern. Recently, several works have been conducted that attempt to develop such a tool for this complicated pattern recognition [Arnold *et al.*, 2009][Samudrala *et al.*, 2009]. These new tools could be easily accommodated by the workflow in the future.

Even though the workflows represented in this chapter did not provide a specific cellular localisation prediction for a protein sequence, the concept of workflow construction could easily be expanded to do so. By utilising Microbase, a loosely-coupled collection of components that together form a distributed computation system, new components or bioinformatic tools could be added to the existing workflows. Microbase allows workflow step reconstruction or rearrangement to suit different study's purposes. For example, tools to predict protein subcellular localisation such as PSORTb could be integrated to the workflow to aid a more precise identification of protein locations. The workflow developed in this study was not designed to predicted the exact protein subcellular location but instead focus on the identification of a general extracytoplasmic location including transmembrane, cell surface and secreted proteins.

The application use case illustrated the use of the identification results generated by the extracytoplasmic protein identification and classification workflows and showed how to gain a greater understanding of biological questions relating to extracytoplasmic proteins. In the case of the *Bacillus*' proteome analysis, the approach used was capable of detecting several protein sequence features known to be specific or abundant in members of the *Bacillus cereus* group including functional domains involved in cell-surface anchoring, amino acid and peptide utilization, antibiotic resistance, and host interactions (e.g adhesion, colonisation, protein interactions, pathogenesis) [Han *et al.*, 2006].

4.5 Conclusions

With the current growth of public sequence databases and the speed of genome sequencing, high-throughput prediction methods have become increasingly important. The approach used in this study has demonstrated to fulfil the need of a high-throughput sequence analysis workflow in the post-genomics era. The workflows developed performed well in terms of accuracy and sensitivity for the prediction of extracytoplasmic proteins among archaea, bacteria (both Gram-positive and Gram-negative), and microbial eukaryotes. The workflow allows an automatic classification of 981,672 putative extracytoplasmic proteins across 867 microorganisms into appropriate classes with respect to the presence of known targeting or anchoring features. The end results from the workflows described in this chapter, together with the results of the protein domain recognition workflow implemented using the high-throughput computational framework described in the previous chapter have provided valuable outcomes in terms of biological meaning as shown by the analysis of the *Bacillus* proteome data set. Several domains dominant in the cereus species are known to facilitate the microbe's ability to thrive in animal host environments or cause disease in the host body. For instance, the chitin-binding domain and collagen-binding proteins might be specific to microbes that interact with insects or vertebrates, respectively. The NEAr transporter, a heam-binding iron uptake, is abundant in the cereus group. The PapR domain regulates various virulence factors, and is only found across the pathogenic cereus group. Moreover, the results from the *Bacillus*' extracytoplasmic proteome analysis indicate several conserved regions of unknown function that might be important in the *Bacillus*-host interaction.

Chapter 5

Microorganism-Habitat Annotation

5.1 Introduction

In order to allow association analyses between protein domains or families and microorganisms thriving in a given habitat(s), information describing the isolation source or known habitat of an organisms is required. This information then enables a comparison of genomes from organisms adapted to different ecological niches to be performed. However, obtaining such information is a significant challenge due to the lack of a well-organised resource containing habitat or isolation source information for the microorganisms whose genome sequence data have been made available. Typically, only limited and patchy information of microorganisms is accessible via major public genome data centre such as the NCBI and GOLD databases (see section 2.5). Moreover, the availability of information relating to the habitat of a microorganism is frequently under-specified. Nevertheless, there is no data source for habitat-microorganism information allowing programmatic access; it is implausible to obtain habitat annotation computationally when large numbers of genome sequences are available to be studied.

Hence, there is a need to assemble the microbe-habitat information in an ad hoc fashion from various sources independently from the genome sequence data to be analysed [Ahmed, 2009][Field *et al.*, 2008]. Due to the large numbers of taxa for which genome sequence data are available, and their increase on daily basis, there is an urgent need to be able to describe the habitat for each taxon in an automated and consistent fashion [Hirschman *et al.*, 2008].

Published literature was considered as a primary resource to fulfil this task and were suggested as the most detailed resources for information relating organisms' niches [Hirschman *et al.*, 2008]. This type of information is rarely found in the abstract of publications, necessitating a full-text search.

Several examples of successful uses of text-mining in biomedical research (e.g. [Groth *et al.*, 2008] [van Haagen *et al.*, 2009] [Rzhetsky *et al.*, 2008]) prompted us to consider the text-mining approach as useful for extracting habitats or isolation sources of microorganisms used in our genome-wide study.

In addition, habitat terms must also be classified with a set of controlled vocabularies representing widely used terms referring to a generic type of ecological niche, as well as more specialised terms where necessary. A stable set of controlled vocabularies referring microorganisms' niches does not yet exist. Several projects are working toward this goal, including Environment Ontology [1] (EnvO) and Habitat-Lite [Hirschman *et al.*, 2008] [2]. However, there is currently no standard set of concepts for the classification of low level habitat terms into high level classes that differentiate properties of habitats either geographically or anatomically.

Since this project aimed to identify gene-encoding protein sequences that are specific to mucosal-lined niches, one habitat of particular interest for this project is human mucosa, and more generally the mucosal surfaces of animals. In particular for organism-associated habitats, there is a need for hierarchical sub-classifications, providing appropriated anatomical differentiation. A detailed anatomical habitat classification can be constructed by integrating high-level classes of an anatomy ontology or controlled vocabularies [Hirschman *et al.*, 2008]. To our knowledge, no one has extended a habitat ontology to describe host parts such as human parts or organs or tissue (like mucosa) (e.g. [Baldock and Burger, 2005]) or any other host parts and environmental habitat such as lakes, soils etc. [von Mering *et al.*, 2007]. The work described in this chapter addresses the development of a suitable ontology, covering both organism-associated and environmental types of habitat to some level of detail.

5.1.1 Objectives

The aims of the work presented in this chapter were to: 1) develop a specific text-mining tool to extract a set of microorganism-habitat attributes from the vast amount of available literature; 2) construct a set of common terms used to refer to habitats of microorganisms.

For the purpose of differentiating mucosa-associated habitat terms from others, a project-specific habitat ontology was constructed in order to: 1) provide consistent high level guidelines for the habitat classification, taking into account geographical and anatomical characteristics; 2) allow the

[1] `http://www.environmentontology.org/`, accessed 21st April 2010.

[2] `http://gensc.org/gc_wiki/index.php/Habitat-Lite`, accessed accessed 21st April 2010.

utilisation of a rule-based approach to facilitate automatic classification of the habitat terms obtained from the text-mining.

This chapter includes:

- the manual microbe-habitat annotations available from public genome resources;
- the development of a text-mining tool to extract microorganism-habitat information;
- the performance of the text-mining tool developed;
- the development of a project-specific habitat ontology;
- the classification of habitat terms using ontological reasoning and results obtained.

5.1.2 Terminology

The following terminology is used throughout this chapter.

- Habitat: from Hirschman et al. 2008 [Hirschman *et al.*, 2008] “the place or environment where an organism naturally or normally lives and grows”.
- Isolation source: a natural source where microorganisms were isolated from, including a anatomical regions infected by pathogenic microbes and body secretions containing either symbiotic or pathogenic strains.
- Mucosa-lined surface: vertebrate epithelial surfaces covered by mucous membrane. Often found on various body cavities that are exposed to the external environment e.g. intestinal lumen, oral cavity, genital area. (See Figure 2.3 for human mucosa surfaces)
- Mucosa-associated microorganisms: a microorganism is labelled as mucosa-associated if there is an evidence showing that at least one of these statements is true: they grow on or colonise mucous membranes; a mucosal environment is a part of their life cycle; they are pathogenic on or through mucosal surfaces; they were isolated from a mucosa-associated area. Some organisms may have multiple habitat such as *Vibrio spp.* (sea water and human digestive tract).

5.2 Materials and methods

5.2.1 Manual microbe-habitat annotation using public databases

Genome sequence information or information about organisms can be retrieved from two well-known genome databases, the NCBI [3] and the GOLD databases [4]. These public genome databases provide relatively similar information relating to genomes and the phenotype of source organisms. However, NCBI genome information is provided in two separate file formats for prokaryotic and eukaryotic genomes, respectively; slightly different kind of information provided in each file type. Therefore, for practical reasons, the genome information from the GOLD database (downloaded 22th october 2009) was used in this project since the information is provided in a homogenous form for both prokaryotic and eukaryotic genomes. The homogenous form of the data from the GOLD database facilitates data parsing and integration with other data sets in the project. Fields considered to determine habitats or isolation sources of a microorganism are ‘isolation site’, ‘body sample site or subsite’, ‘body product’ and ‘disease’. The ‘habitat’ field was also considered even though this field is often empty. The ‘isolation site’ field provides a relatively detailed description of genome isolation sources. The ‘body sample site or subsite’ and ‘body products’ contains useful anatomical information and secretion products of animal hosts. The ‘disease’ field sometimes contains specific terms related to a disease which can be used to infer isolation sources where pathogenic species or strains thrive or colonise. This field often provides an indication of the ability of certain microbes to infect vertebrate hosts through mucosa-lined surfaces. These fields were used as a primary source of knowledge for habitat information of a microorganism included in this project.

In this study, terms considered as mucosa-associated and stated in the derived GOLD genome information are listed as follows.

- Mucosa-related digestive parts: gingival, dental, periodontal, oral, mouth, intestinal, rumen, caecum, appendix, gastric, enteric, saliva, fecal, feces, periodontitis, gastroenteritis, colitis, diarrhea, food poisoning, botulism, cholera, dysenteria, thyphoid.
- Urogenital parts: urogenital, vaginal, genital, urinary, bladder, gonorrhea, trichomoniasis
- Respiratory parts: airway, respiratory, pulmonary, sinusitis, pneumonia, tuberculosis, anthrax
- Other parts: eye, ear, mammary gland, ocular, otitis.

[3] ftp://ftp.ncbi.nih.gov/genomes/genomeprj, accessed 21st April 2010

[4] http://www.genomesonline.org/, accessed 21st April 2010

If any of the GOLD fields listed above contained any of the project mucosa-associated terms, the microorganism linked with those terms was assigned as a mucosa-associated microorganism. Otherwise, microorganisms were assigned to other ecological niches, based on the information provided in the relevant fields. Other habitats often stated in the GOLD fields were, for instance, soil, plant, hot spring, sea, sediment etc.

5.2.2 Text-mining to extract microbe-habitat information

The goal of using text-mining techniques is to efficiently discover microorganism-habitat pairs by automatic integration and analysis of the literature, rather than a conservative approach of manually searching and reading through the text. It is anticipated that the text-mining approach will lead to the discovery of many true positive attributes and to enrich existing habitat annotation [Cohen and Hunter, 2008].

The first step towards that goal was for the text-mining tool to recognise key entities such as organism scientific names and terms referring to habitat or isolation source of microorganisms. The next step was to extract organism-habitat relation pairs from published literature. This task presents new challenges for text-mining: there is no prior standard annotated corpora to serve as training data for machine learning algorithms, or to provide a gold standard for evaluation; and information of interest frequently appears in the main text rather than the abstract of the publications [Levow, 2010, pers. comm.].

As a result, new corpus materials were developed with an aim of training and evaluating these concepts of interest based on the annotation of full text in the literature [Levow, 2010, pers. comm.]. The accelerated annotation (Acela) interface [Tsuruoka *et al.*, 2008] was used for interactive annotations of text for both microorganisms and habitat entities, as well as organism-habitat relation pairs. This interface allows for interactive and iterative training of a machine learning classifier to recognise a specific entity class or concept. This tool has proven most useful for concepts which appear least frequently in the corpus by directing attention to the relatively few sentences in which most concepts are likely to appear [Levow, 2010, pers. comm.].

Corpus creation and annotation

For the microorganism-habitat corpus, two classes of entity including microorganisms and habitat or isolation source, were annotated for the key entity recognition step. Manual exploration was carried out of various patterns of explicit habitat-associated sentences in numerous publications for taxa-habitat pairs of both host- and non host-associated microorganisms. The list of manual taxa-habitat annotations containing approximately 57 taxa-habitat pairs from 20 publications (shown in Appendix B) was initially used as a seed for the initial training of the text-mining system. Annotator instances of the Acela for the two classes were created. A new set of sentence examples in which interesting terms were labelled based on the initial state of the classifier were presented through the Acela interface for human validation or correction. Annotation results were used iteratively to train the classifier interactively. The annotation was performed via the interface to label instances until an estimated coverage over 99% was achieved [Levow, 2010, pers. comm.].

The criteria used by an expert to train the machine learning system to extract terms referring to the scientific names of microorganisms and terms inferring the isolation source or habitat of a microbe are described below.

Microorganism entity

The text mining system was able to tag organism terms where sentences contained sufficient information to identify the organism:

- Organisms could be tagged where the scientific name of microorganisms was specified to at least the Genus level. Microorganisms include bacteria, archaea, microbial eukaryotes. Specie, strain, and serovar entries are also tagged if they are present. Examples of terms tagged for this entity are *E. coli*, *Campylobacter spp.*, and *Trichomonas vaginalis*.
- Organism terms could also be tagged where sentences also contain habitat or isolation information, in addition to the organism. For example, the microorganism name was tagged for the sentence: '***Bacteroides salyersae sp. nov.*** isolated from clinical specimens of human intestinal origin'.

Habitat entity

The following types of sentence may be utilised by the text-mining system for the annotation of terms relating to the habitat or isolation source of a microbe:

- Context-related, or a reference to a habitat or isolation source of an organism e.g. 'Isolation and distribution of bartonellae in wild **rodents** in Japan'.

- Habitat-related terms used as an adjective describing an organism e.g. '**oral** Campylobacter', '**rumen** bacteria', '**rodent**-associated Bartonella febrile illness'.

- Terms representing isolation source of an organism e.g. 'The cases included a **breast abscess** caused by *Campylobacter rectus* and a non-group A beta-hemolytic Streptococcus in a patient with lymphoma, a **liver abscess** caused by *Campylobacter curvus* and an alpha-hemolytic streptococcus in a patient with complicated ovarian cancer, and a postobstructive **bronchial abscess** caused by *C. curvus* and group C beta-hemolytic *Streptococcus constellatus* in a patient with lung cancer'.

- Terms not associated directly with disease e.g. 'respiratory tract infection', 'diarrhea', and 'periodontal disease'.

- If the habitats or isolation sources are another organism species, the terms were tagged with their common name or Genus name without the species names e.g. 'microbeA was isolated from ***Apodemus spp.***'.

Recognition approaches for the text-mining

In order for the text-mining system to recognise terms representing the interest entities, two approaches were employed for the recognition approach: a dictionary-based approach; and a hybrid machine learning approach with dictionary information. The work presented in this section was designed and conducted in collaboration with text-mining experts from the National Centre for Text Mining (NaCTeM[5]) at the University of Manchester (see Figure 5.1).

Terminological resources

For both entities of interest, lexical resources were constructed based on a combination of curated domain ontologies and a list of terms from existing resources. Resources for microorganism scientific names were obtained from the NCBI taxonomy[6] and the 'List of Prokaryotic names with Standing in Nomenclature' (LPSN)[7]. All scientific names from these resources were extracted. The names were then converted into standardised forms covering different typical variability for the term. For example, the tags representing taxonomic levels, such as 'subsp.', 'str.', 'strain', were removed from

[5] http://www.nactem.ac.uk/, accessed 10th December 2010

[6] http://ncbi.nlm.nih.gov/taxonomy, accessed accessed 21st April 2010

[7] http://www.bacterio.net, accessed accessed 21st April 2010

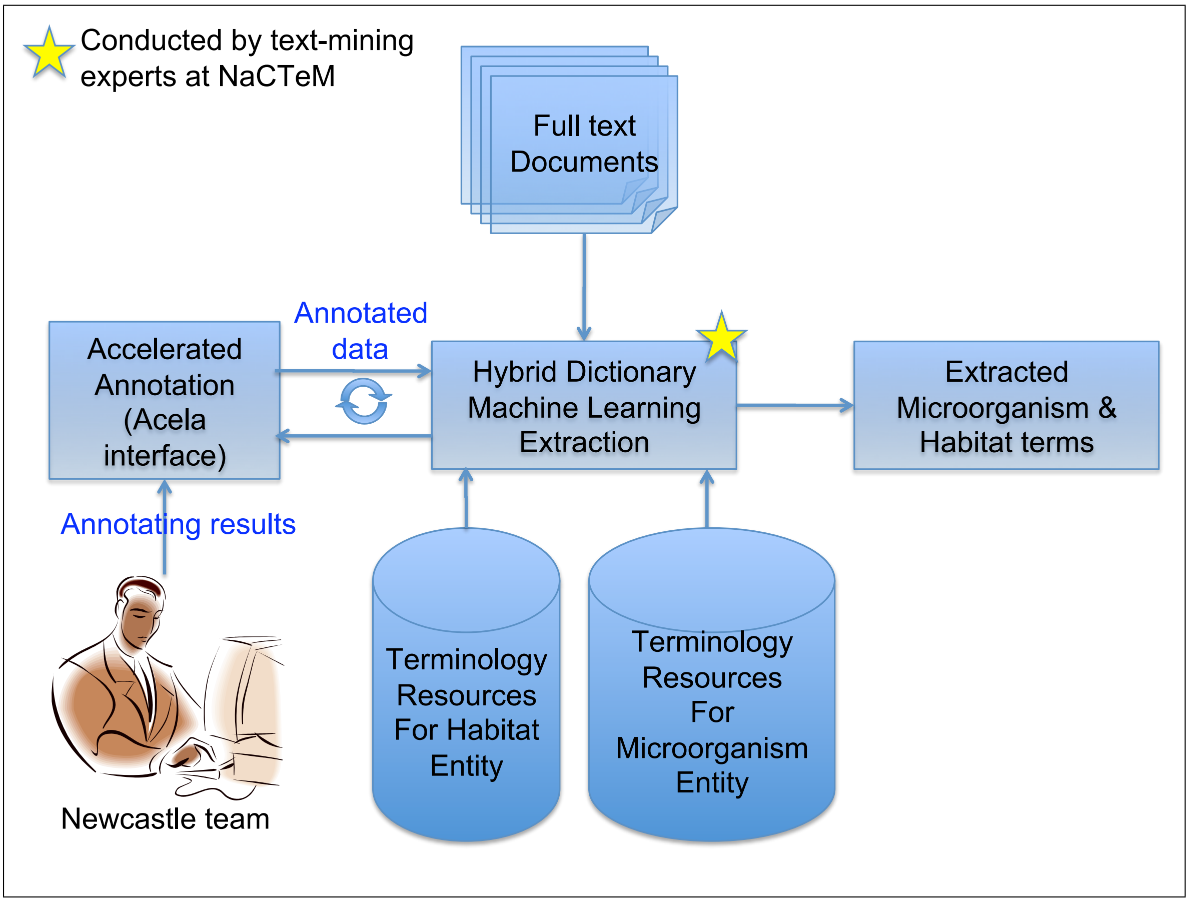

Conducted by text-mining experts at NaCTeM
Full text Documents
Annotated data
Accelerated Annotation (Acela interface)
Hybrid Dictionary Machine Learning Extraction
Extracted Microorganism & Habitat terms
Annotating results
Terminology Resources For Habitat Entity
Terminology Resources For Microorganism Entity
Newcastle team

Figure 5.1: Flowchart summarising the text-mining system developed for extracting microbe-habitat attributes from literature. This work was done in collaboration with text-mining experts from the National Centre for Text Mining (NaCTeM) at the University of Manchester. The yellow star denotes the step conducted by the text-mining experts. The other aspected depicted in the diagram were performed as a collaboration between the author and the text-mining experts. The author provided initial seed documents and annotated the results from the text-mining system through the Accelerated annotation (Acela) interface. The annotated data was then used to train the system. The annotation, extraction and training steps were iteratively cycled until the system performance reached a satisfactory level. The technique for the extraction (recognition) of terms in microorganism and habitat entities is a hybrid machine learning approach with dictionary information.

the name. As a result, the term list for microorganism entity comprises 52,715 entries for 12,256 distinct organisms.

For the habitat entity, 135 terms referring to habitats/isolation sources from the GOLD database were used as the main resource (accessed 17 August 2008). This set of habitat terms was further enhanced with 12,0668 entries containing names of animals, organs and body parts extracted from the UMLS Metathesaurus [8].

5.2.3 Classifying habitat term-based to knowledge-based

The previous section describes a collaborative work with NaCTeM regarding the development of a text-mining tool to extract microbe-habitat information. This section represents the work performed by the author of this thesis upon the developing of a habitat ontology.

In order to standardise the text-mined terms referring to habitats into a set of controlled vocabularies capturing generic types of microorganisms' niches, an ontology containing terms representing generic classes of microorganisms' habitats was developed. These terms permit high level differentiation between physical and chemical properties of habitats, either geographically or anatomically. These generic classes may have a relationship indicating parenthood and childhood between the terms. One term can have multiple subclasses. For example, 'Aquatic' and 'Terrestrial' are two high-level generic terms referring to two very different ecological properties. The former term represents a water-related space, while the latter concerns area relating to land or earth. 'Aquatic' can be sub-classified into 'Saline water' and 'Freshwater'. These two subclasses share the properties of a water-based habitat with their parent, but allow divergent sub-properties to be represented; in this case, the presence or absence of salt. Moreover, a subclass can have multiple inferred parent classes. For example, 'RespiratoryPart' class is asserted under 'AnatomicalPart', and have a property of being lined with mucosa. Therefore, the 'RespiratoryPart' class is also has an inferred parent as 'MucosaLining' habitat as the 'MucosaLining' is defined by any organism part that is lined with mucosa.

The high-level habitat terms were carefully selected by considering the information contained in the GOLD fields. These generic terms include some second level terms from the Habitat-Lite ontology, version 0.3 [Hirschman *et al.*, 2008]. Habitat-Lite was the first ontology to establish concepts for describing high-level terms relating a limited set of habitats of organisms. The concepts in Habitat-Lite form a simple hierarchy of a single type of relationship, providing a light-weight set of

[8] http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html, accessed 21st April 2010

terms describing habitats. Several terms were also adopted from the Environment Ontology (EnVo) [Morrison and Field, 2010] which partly covers terms describing habitats.

Apart from providing a standard generic classes of microorganisms' habitats, the ontology developed in this project was also to represent a lexical resource of common terms used to referred to habitats isolation sources of microorganisms. Therefore, these terms were assigned as synonyms or labels of the corresponding sibling generic classes. For example, 'Marine' and 'Sea' are both considered as synonyms of 'Saline water' which is considered as a generic habitat class.

The project-specific ontology was developed in the Web Ontology Language (OWL) using Protege (4.0). Protege is an ontology editor and knowledge-base framework [Rubin *et al.*, 2007]. The reasoning algorithm, Pallet version 1.5.2, was employed as an OWL reasoner to reason and query the information in the ontology.

5.3 Results

5.3.1 Manual microbe-habitat annotation

In this section, the results of microbe-habitat annotation performed manually by using information from the GOLD database is summarised. The results presented in this section were used for computationally identify genotypic features over-represented in a group of microorganisms in relation to their ecological niches (see Chapter 6). The section is divided into two parts:

- Mucosa VS non-mucosa microorganisms annotation: strategies used and the results of classifying microbes into mucosa and non-mucosa associated;
- Comprehensive microbe-habitat classification: an overview of the classification of microorganisms based on the information from the GOLD database.

Mucosa VS non-mucosa microorganisms classification

The focus of this study was primarily to distinguish microorganisms that are able to thrive in a mucosal environment from others. Therefore, to identify if a microbe is mucosa-thriving, the information relating to animal host-associated isolation sources such as body site, body secretion and diseases were the first information to be considered. Free-living microbes were also labelled as 'mucosa-associated' if they cause disease to animal hosts via mucosa surfaces, even though they are known to live in multiple environments. For example, *Vibrio cholera O395* was noted as a cause of pandemic food poisoning but originates from aquatic environments [Nelson *et al.*, 2009]. Microorganisms were classified as ‘non-mucosa-associated’ if there is no information of their isolation sources that can be attributed to mucosa-lined niches. Some microorganisms with ambiguous isolation information in terms of whether it is related to mucosal surfaces, were not assigned to either of the classes. For example, some microorganisms were only provided with disease information such as "Toxemia" or "Septicemia". Additionally, some taxa with unclear or complicated life cycles that involved animal hosts were also not included in the classification process. These organisms include, for instance, the members of *Bacillus cereus* group and *Leishmania spp*.

As a result, 203 out of 867 microorganisms in the GenomePool database (GPDB) were classified as mucosa-associated organisms (see Figure 5.2). This data set contains both allochthonous and autochthonous residents of mucosa-lined niches. The set of non mucosa-associated contains 320 taxa from the GPDB, leaving approximately one third (344) of microorganisms unclassified.

Microbe-habitat classification

In this section, organisms are categorised based on their taxonomic groups and their associations with high-level habitat terms (see Figure 5.3). For taxa considered to be host-associated microorganisms, their roles in the relationships with hosts were assigned where possible. These roles include terms such as ‘pathogen’ and ‘symbiont’. The term ‘pathogen’ is assigned to microorganisms that are known to be free-living as well as known to cause disease when interacting with a host body. The role ‘symbiont’ is assigned to microorganisms that naturally inhabit a host body and are known to stabilise the host physiology.

Approximately 15% (126/867) of the total number of taxa in the GPDB have no information relating their isolation sources in the GOLD database. These taxa were therefore classified into ‘unknown source’ and were not used in the genotype-habitat association analysis.

5.3.2 Microbe-habitat information discovered through a text-mining approach

A text-mining system for the automated discovery of microbe-habitat pairs from available literature was developed. We collaborated with text-mining experts at the NaCTeM with the aim of developing a high-throughput system that would remove the need for manual annotation. Different techniques for the recognition of two different types of entities: microorganisms’ names and habitat terms were investigated. The annotations generated by the system were verified based on the author’s biological

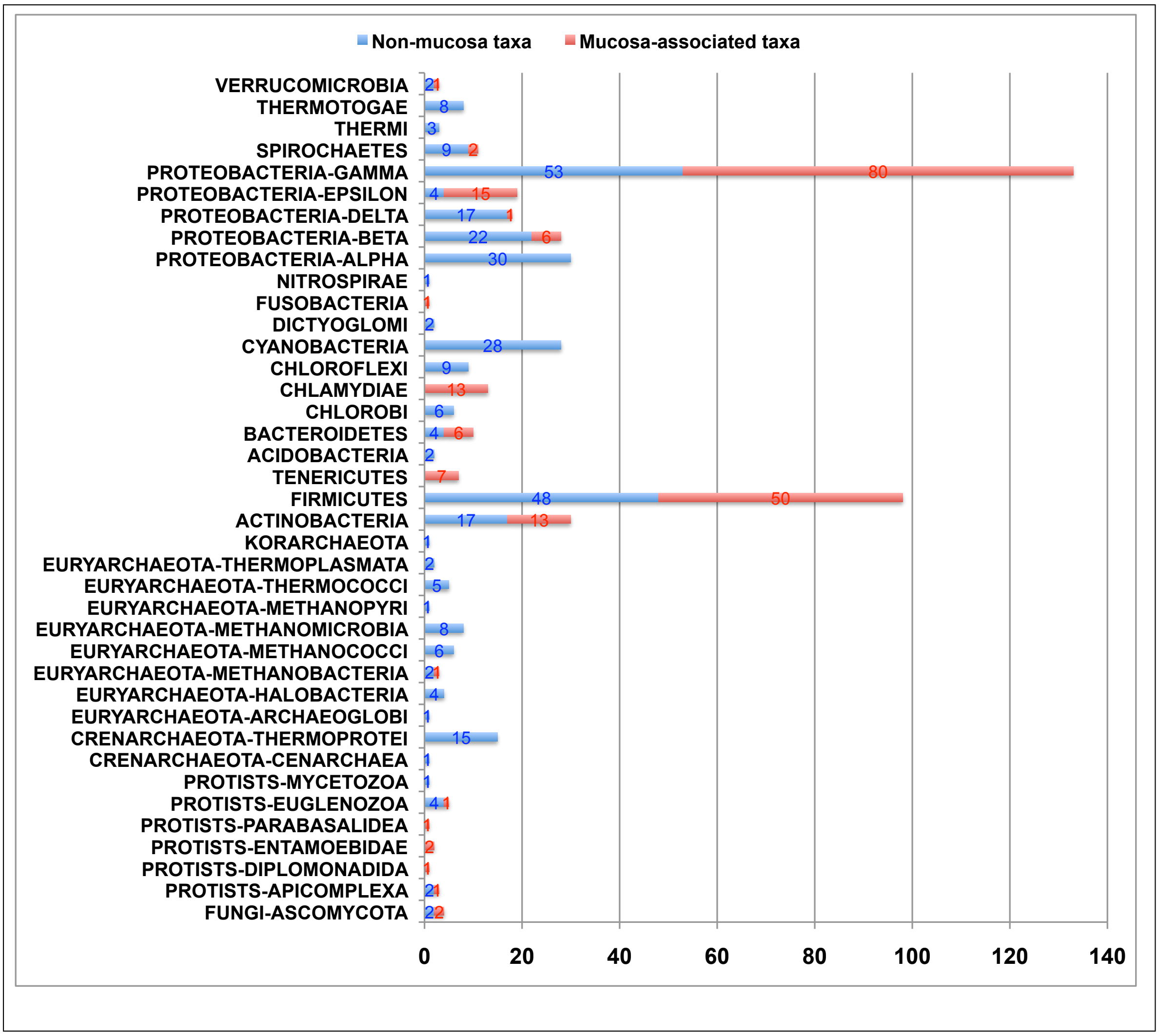


Figure 5.2: Summary of the number and the distribution of taxa included in the mucosa-associated microorganisms classification. The taxa-habitat annotation was performed manually based on the genome information derived from the GOLD database. The X-axis represents the number of microorganisms species or strains. Red bars indicate mucosa-associated taxa, while blue bars represent non mucosa-associated taxa. Mucosa-associated microorganisms are presented in Archaea, Bacteria, Fungi and Protists. The largest proportion of mucosa-associated taxa are in Proteobacteria-Gamma and Firmicutes.

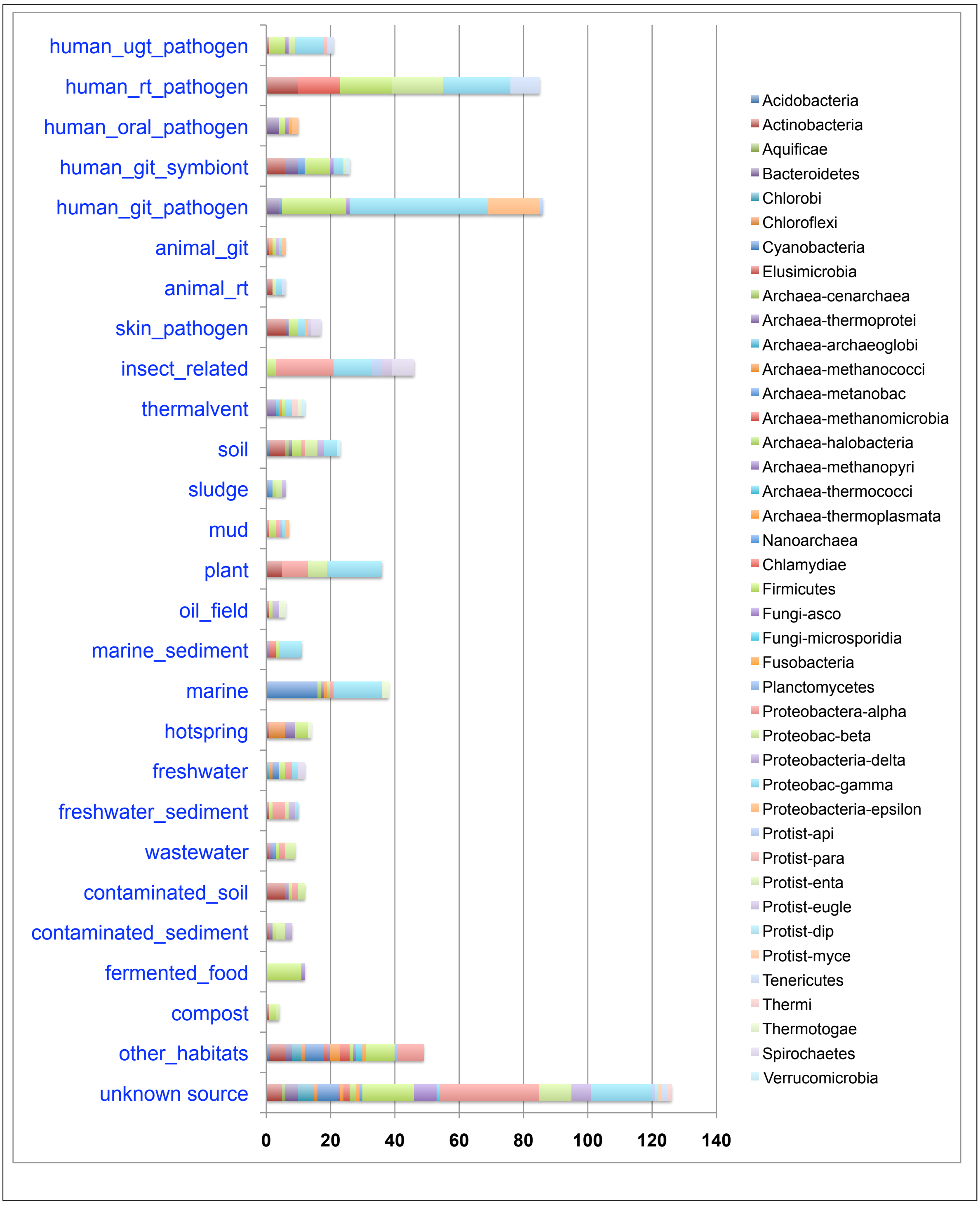


Figure 5.3: Summary of the number of taxa classified by their habitats or isolation sources. The taxa-habitat annotation was performed manually based on the genome information derived from the GOLD database. The X-axis represents the number of microorganisms species/strains isolated from a given habitat or isolation source. A taxa can be annotated with more than one source type if it was found in multiple sources. Mucosa-lined environments are shown in the red square where 'ugt' is urogenital tract; 'rt' is respiratory tract; and 'git' is gastrointestinal tract). For the mucosa-associated environment, host-microbe relationships were also categorised into 'symbiont' and 'pathogen' where possible.

knowledge and the context of sentences extracted. For the organism tagging, the conditional random fields (CRF) approach outperformed the dictionary-based approach (F-measures 63% and 80%, respectively). The CRF technique employed machine-learning strategies [Levow, 2010, pers. comm.]. The result also indicated that the restriction of annotation to organisms within habitat contexts has a significant impact, as proved by the improvement in precision rate yielded by the machine learning setting (data not shown). The habitat recognition was shown to be challenging; none of the techniques returned satisfied results (F-measures ranged between 50-56%) [Levow, 2010, pers. comm.]. The errors made by the system indicate both false negatives and false positives with the same words (e.g. human, water). In particular, there were issues with the generalisation over adjectival forms of the terms referring to habitats, such as 'extraoral' [Levow, 2010, pers. comm.]. Running the system on a dataset of 9,265 full text documents from the Open Access subset of PubMed Central from 2007 resulted in a relatively high degree of false-positive pairs. However, some true positive results can also be obtained and already appear to provide promising outcomes (see Appendix C).

5.3.3 The development of a habitat ontology

The word 'habitat' was defined as an abstract term used to indicate a role of any physical spaces or object that is a place of residence to any living organism. By this definition, any existing object in the universe could be referred to as a habitat of something. In order to produce a sensible and practical habitat ontology to serve the needs of this project, we investigated terms stated in the GOLD database that provide habitat information. Terms commonly used to describe the same or similar ecological niches with respect to their general physical and chemical properties were grouped together. Generic habitat terms, or so called the entity classes, were created based on the consideration of the common property of each group and the usage of the classes in the later stages of the project. In other words, these generic terms were selected with our interest and the project's research questions in mind.

As a result, the habitat ontology developed in this project expanded the existing Habitat-Lite ontology (V.0.3) by integrating other terms that can also be described as a place where microorganisms may thrive. The main focus was to cover terms denoting host anatomical-related niches in details in order to support the requirements of this project; namely, the inference of mucosa-associated environments where appropriate. Two other main entities were added to the Habitat-Lite. These entities are: a basic ontology describing animal anatomical location and types of organism hosts. The structure and entity class names of the anatomical ontology developed as part of this project was inspired by, and adopted terms from other existing anatomical ontologies, including the Foundational Model of

Anatomy (FMA)[9] Ontology, and Common Anatomy Reference Ontology (CARO)[10]. The adapted anatomy entity, designed to facilitate the later analysis stages of this project, focuses mainly on vertebrates and in particular human anatomy. The anatomy entity is composed of classes describing 8 major human anatomical parts and 2 other classes denoting types of the surface epithelium; mucosa and non-mucosa (see Figure 5.4). The terms representing anatomical parts, commonly found as a niche and lined with a mucous membrane, can be inferred to be the 'MucosaLining' class. As a result, the ontology can be queried using terms found from the GOLD database or from information mined from the literature. Terms will be automatically classified into appropriated habitat generic classes if the terms are assigned as labels of an appropriate entity class. For example, 'RespiratoryPart' entity class can be inferred as 'MucosaLining' habitat, a term assigned as labels of the 'RespiratoryPart' class such as 'Lungs' is also therefore referred as the 'MucosaLining' habitat.

Another entity present in the habitat ontology developed represents types of organism hosts. The concept classes composing the organism type entity were selected on the basis of types of organisms commonly reported as hosts. The relationships between each class were chosen with the NCBI taxonomic classification in mind. The entity represented four main high-level classes: 'Animal', 'Plant', 'Protist' and 'RoledOrganism' (see Figure 5.5). The 'RoledOrganism' class was introduced in order to allow the organism to be presented with a role 'host' which can be then inferred as 'OrganismAssociatedHabitat' in the habitat entity. Every subclasses in the 'RoledOrganism' class are still organisms, even though conceptually has a role as host. The 'OrganismAssociatedHabitat' class can be then defined by any organism that has 'host' role.

As a result, the project-specific ontology represents knowledge of microorganisms' habitats of both environmental and host-anatomical niches. The inferred model of the habitat entity is shown in Figure 5.6.

5.4 Discussion

5.4.1 Microorganism-habitat information in public sources

Manual organism-habitat annotations were performed in order to generate a data set for training the machine-learning text mining algorithm. One interesting point that arose from this manual annotation was that detailed information of the sources or niches for particular microorganisms for which the

[9] http://obo.svn.sourceforge.net/viewvc/obo/fma-conversion/trunk/fma2_obo.obo, accessed 21st April 2010

[10] http://obo.cvs.sourceforge.net/viewvc/obo/obo/ontology/anatomy/caro/caro.obo, accessed 21st April 2010

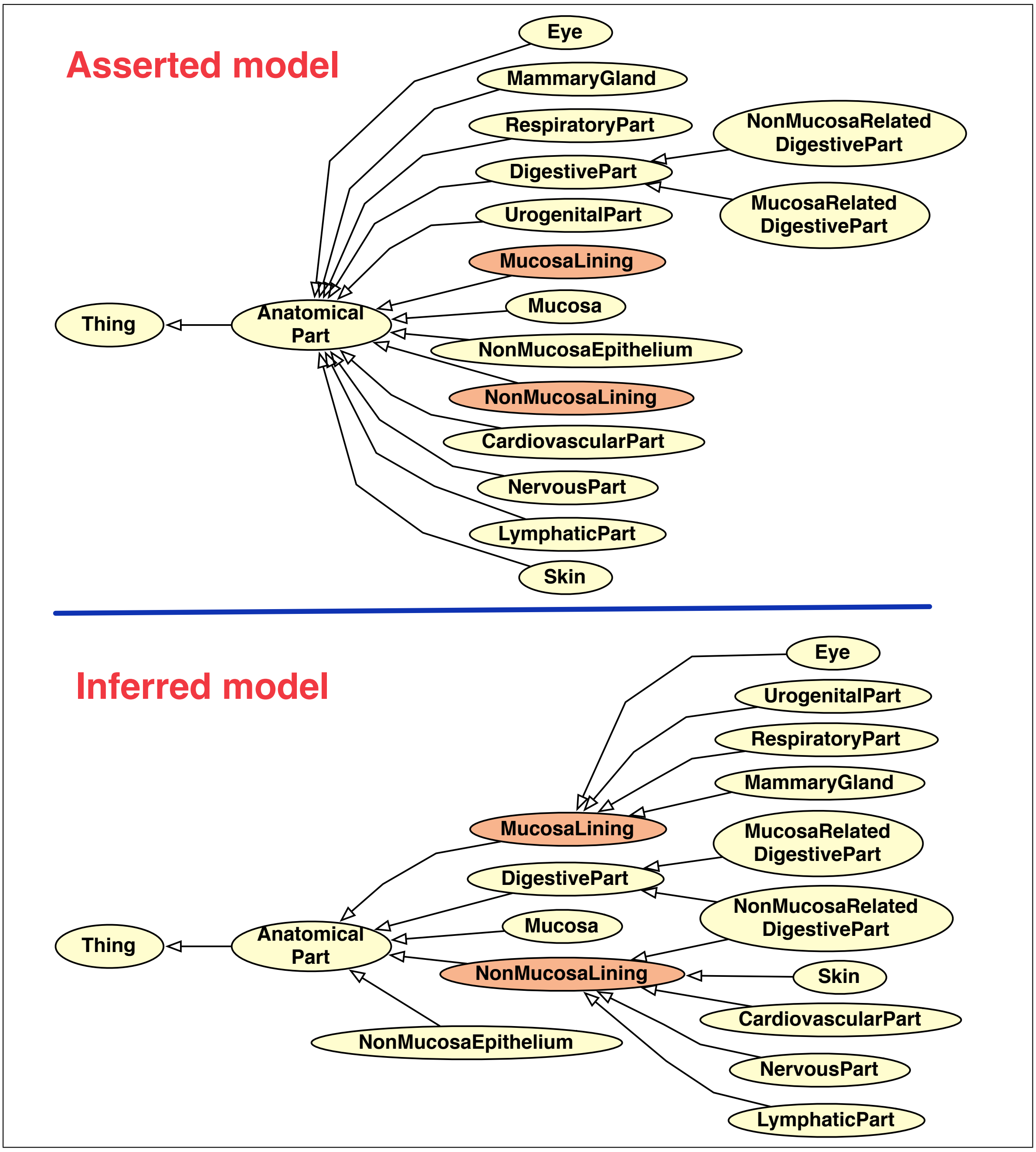


Figure 5.4: The anatomy entity represented in the habitat ontology developed for this project. Both asserted and inferred models of the anatomy entity are shown. The asserted model represents classes and their relations before the application of a reasoner. The inferred model illustrates the classes and their relations after a logical reasoner has processed the asserted model. Dark yellow nodes represent defined classes, i.e. classes with constraints associated with them in order to facilitate automated inference. Executing a reasoner over the asserted model to produce a new, inferred hierarchy has certain advantages. For example, the reasoner has inferred that concepts such as 'Eye' and 'UrogenitalPart' are mucosa-lined, whereas concepts such as 'Skin' are not mucosa-lined. Automated inference techniques allow the asserted model to be checked for consistency, as well as suggesting links between concepts that may not have been explicitly added by the author.

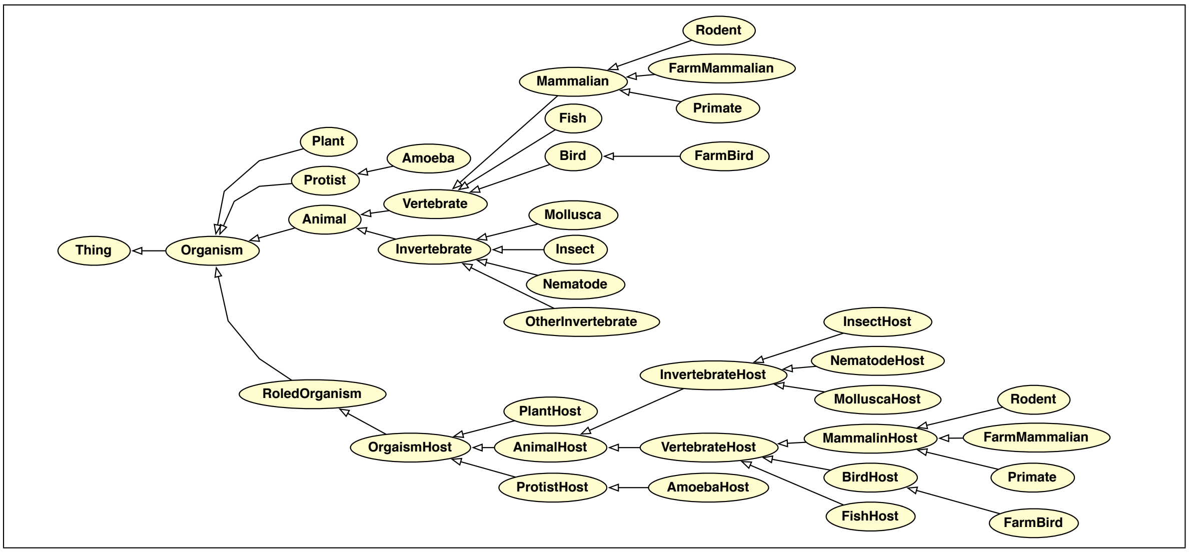


Figure 5.5: The organism entity represented in the habitat ontology developed for this project. The diagram represents asserted models of the entity. The inferred model is not shown as no inference was made. Dark yellow nodes represent defined classes. See Figure 5.4 for the meanings of asserted and inferred models.

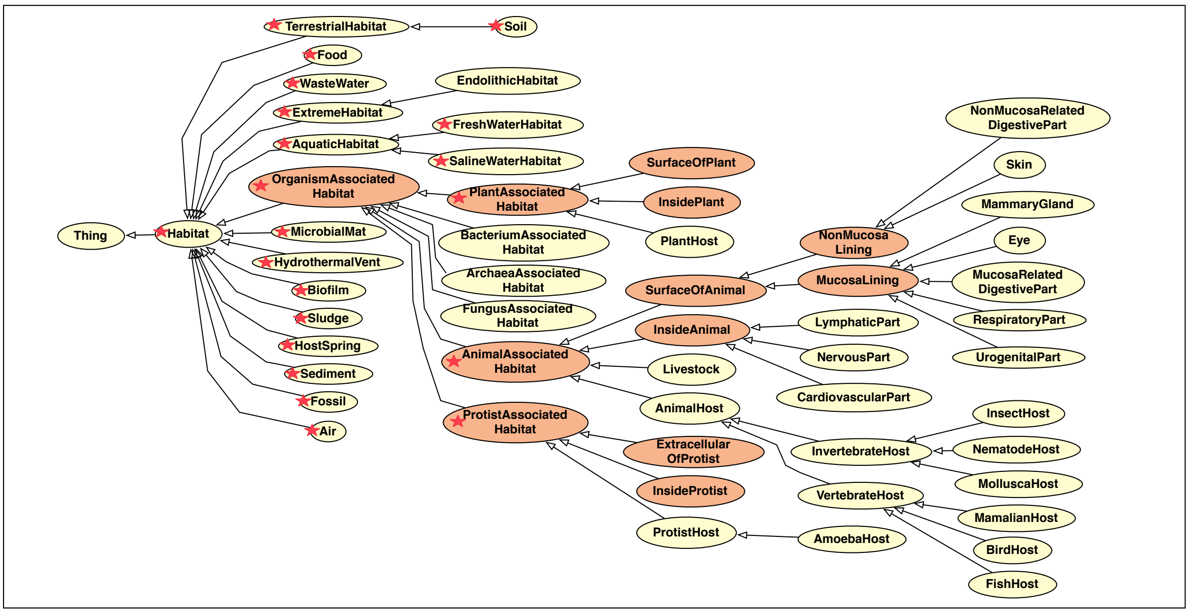


Figure 5.6: The habitat ontology developed for this project. The inferred relationships between high-level generic habitat classes are shown. Dark yellow nodes represent defined classes. Each class represents an abstract microorganism habitat, rather than a specific habitat; for example, the class ‘UrogenitalPart’ would represent terms referring habitats such as ‘bladder’, ‘urethra’ and ‘uterus’. Some class names were obtained from the Habitat-lite ontology V.0.3 and are marked with ‘*’. See Figure 5.4 for the meanings of asserted and inferred models.

genome sequence is available, is often poorly defined in the literature if mentioned at all. Therefore, the information about the environment or niche where an organism resides is generally obtained from fundamental knowledge of the ecological study of that organism, but may not necessarily correspond to a specific organism strain from which the genome sequence data was derived.

The information about a microorganisms' habitat provided in public databases does not cover all those organisms for which complete genome sequence data is available. Moreover, the 'habitat' field in the GOLD database is frequently empty. Due to the lack of direct habitat terms specifying mucosal environments, manual annotation of microbe-habitat data from disease fields were used to specify whether a microbe is mucosa-associated. Computational text-mining is an approach that can potentially overcome this problem, allowing more terms referring to habitats to be acquired from the literature.

The information about the habitats of organisms available in the public database is human microbe-centric and does not provide as much information for other animal or environmental microbes.

5.4.2 Text-mining for the microbe-habitat annotation

The recognition of habitat terms by the text-mining system was shown to be particularly challenging. A very wide range of class definitions including animal, anatomy, and environment terms of both noun and adjective forms are present as habitats or isolation sources [Levow, 2010, pers. comm.].

Further refinements to the text mining tool could be made. For example, the system could be trained with a greater number of documents in order to further tune the system, improving both accuracy and sensitivity of detections. Removing terms that are not likely to be habitat-related such as 'process', 'test', and 'helix' (see Appendix C) from the dictionary used as a resource of habitat terms might also reduce the number of false positive values. The addition of more specific terms that are not yet present in the dictionary, such as 'mucosa', 'gut', and 'fallopian-tube', should improve the ability of the tool to extract specific habitat terms.

5.4.3 The habitat ontology as an aid to inferring knowledge

The use of an ontology facilitates the representation of knowledge into concept classes and rich relationships between the concepts. This formal representation also allows machine readable and programmatic access, permitting concepts to be automatically reasoned over based on their properties within the domain. Therefore, new knowledge and relationships can be inferred based on the asserted knowledge provided by the developer. The framework enables the expansion of concepts representing knowledge. New terms can be easily added to the appropriate generic or concept class(es). Similarly, new concepts can be added into the ontology to suit a research interest. Even though the ontology has not been used practically with a large term set. it has been tested with a small scale and yielded a relatively good results (data not shown).

In order to classify habitat terms not yet presented in the ontology, these new terms must be added into the domain. The method to complement these terms to the domain is subjected to an on-going investigation.

5.5 Conclusions

This study highlights the lack of a well-structured, coherent source of data relating to microbe-habitat information that is amenable to automated querying techniques. The information available in the widely used public genome databases, such as GOLD and NCBI, is growing at a much slower rate than the exponentially increasing number of complete genome sequences. There is therefore a lack of verifiable metadata about many genome sequences. This metadata, carrying isolation source characteristics or other important characteristics such as the phenotypes of the source organisms, is an important key to allow the association of genome content to habitats. The work presented in this chapter has shown how this limitation can be overcome to some extent, through the text-mining of free-text from multiple publications into a structured ontology. The results from the genotype-habitat associations (presented in the next chapter) provide a better understanding of the genotypic features that are involved in the survival of a microbe in particular ecological niches.

The use of text-mining techniques to obtain microbe-habitat annotations is one of the first systematic, automated methods developed to obtain microorganism-habitat information scattered throughout the literature. The investigation and experimentation with a text-mining system for extracting this information provides a proof-of-principle that could become a feasible means of obtaining habitat information with future work. The work to date has investigated the application of entity recognition techniques to support the automatic extraction of microorganism names and terms referring their habitats. The annotation corpora for these novel key entities were created through the use of the Acela interface, which reduces the complexities involved in manual annotation.

A prototype ontology for organising and reasoning over the habitats of microorganisms has been developed. The ontology represented generic classes for the habitat of both environmental- and anatomical-related habitats. Each of these habitat entity classes holds a set of terms or controlled vocabularies referring to that class providing a lexical resource for terms referring microorganisms' habitat. This is the lexical resource for microorganisms' habitat terms that is back-ended with an ontology that is capable of inferring a mucosa-lined habitat based on which body part of vertebrate host was selected.

In summary, this work has contributed to facilitating large-scale comparative genomics studies where the ecological niches of microbes are the key focus of the research question.

Chapter 6

In silico Identification and Characterisation of Mucosa-associated Proteins

6.1 Introduction

The ever-increasing availability of genomics data provides an opportunity to perform detailed comparative genomics studies. A comparison of both multiple DNA and protein sequences can reveal potentially interesting genotypic differences among species [Boekhorst *et al.*, 2006][Cornell *et al.*, 2007] or between the microbial communities that inhabit different environments [O'Sullivan *et al.*, 2009] [Kurokawa *et al.*, 2007]. Identifying sets of protein-encoding genes correlated with particular niches can lead to a better understanding of the underlying molecular functions that facilitate the survival of microbes in different ecological context.

Environmental properties vary in different microbial niches. Upon entering a new environment, microbes encounter multiple ecological forces that drive natural selection to allowing them to adapt to the new environment [Lin *et al.*, 2002][Bellgard *et al.*, 2009]. In order to survive, microorganisms may modulate their patterns of gene expression to adapt rapidly to their surroundings, tolerate various external stresses as well as acquire energy and nutrients from a suitable source [Peterson, 2002] [Houot *et al.*, 2010] [Rosenbach *et al.*, 2010] [Dietrich *et al.*, 2003]. The longer term of microorganisms' adaptation strategy to thrive in an ecological niche can occur by altering the genome complement through a number of evolutionary events including gene loss, gene family expansion, lateral or horizontal gene transfer, and mutation [Bellgard *et al.*, 2009] [Ren and Paulsen, 2005].

The resulting microbial community will attempt to successfully adapt to the environment by altering both patterns of regulation of its existing gene repertoire and by modifying its genetic complement. For example, a mucosa microbial community has particular requirements for optimal fitness. These organisms require appropriate cell surface components to attach to the slippery mucous membrane and to avoid rapid wash outs associated with these environments. It is also vital to have a collection of enzymes that allow the use of available substrates as a source of energy and nutrition. Evasion machineries are also required to elude host macrophages and other immune responses [Ley *et al.*, 2006]. Proteins that perform these key functions and biological processes facilitate the ability to competitively thrive in that environment. These protein-coding sequences are likely to be conserved by subsequent generations and adopted by new inhabitants [Xu *et al.*, 2007]. The adoption of the key genotypic features from one microbe to others that live in the same space or habitat and are not from parent to offspring is known as horizontal or lateral gene transfer [Dutta and Pan, 2002][Keeling and Palmer, 2008]. This type of evolutionary event occurs due to selective pressures present in a given ecological condition and can contribute to the ability of microbes to adapt and evolve to survive [Bellgard *et al.*, 2009] [Guénola *et al.*, 2006] [Salyers *et al.*, 2004].

A tremendous number of microorganisms are known to naturally inhabit vertebrate mucosa surfaces such as the gastrointestinal and urogenital tracts. Firmicutes and Bacteroidetes comprise the majority of the microbiome in the human gastrointestinal tract [Rajilić-Stojanović *et al.*, 2007] [Ahmed *et al.*, 2007] [Wang *et al.*, 2003]. These bacteria normally have a mutualistic relationship with the host body include nutrient processing, vitamin synthesis and development of a functional immune system [Turnbaugh *et al.*, 2007]. For example, *Bacteroides thetaiotamicron* metabolise and import indigestible dietary polysaccharides and provide short-chain fatty acids absorbable by the host [Bäckhed *et al.*, 2005][Flint *et al.*, 2008]. Several key elements allowing microbes to successfully thrive on the host-mucosa surfaces have been revealed in the last decade primarily in individual organisms [Acheson and Luccioli, 2004]. For example, the starch utilisation system (sus) was discovered in *B. thetaiotamicron*. The *sus* comprises enzymes and transporters involved in the metabolism of indigestible carbohydrates passed to our distal intestine [Martens *et al.*, 2009].

In this chapter, the distribution of known protein signatures (protein families or motifs) in the available complete microbial genome sequences for which their isolation sources are known was investigated. The investigation was performed in order to identify the protein attributes that are significantly co-occur or are expanded among the known mucosa-thriving microbes. The analysis of their potential functionality and involvement in mucosa-microbe interactions were also carried out.

A genome-wide analysis was performed to identify microbial proteins that have important molecular functions at the host-microbe interface. This analysis involved 3,021,490 protein sequences derived from 867 complete microbial genomes across the domains of cellular life. In this chapter, the ability of microbes to thrive in a mucosal environment was examined in relation to the available functional genomics data. The data generated from the project-specific workflows was further analysed by combining with the microorganism-habitat annotations. The integration analysis results are presented in this chapter.

The chapter is divided into two parts. The first part investigates the functional analysis of the protein domains that are statistically associated with mucosa-related life style of microorganisms. The second part concerns the identification of protein families shared among mucosa-thriving microbes.

6.2 Materials and methods

Comparative genomics was employed to identify candidate proteins that are likely to allow microbes to colonise and thrive in vertebrate mucosal environments. Two approaches were used to determine protein elements specific to mucosa-associated microorganisms. In the first approach, statistical analyses (i.e, association analysis, significance calculation) were performed to associate protein domains with a set of known mucosal microorganisms. The second approach involved clustering extracytoplasmic protein sequences based on their sequence similarities. Clusters containing proteins that were considered to be mucosa-associated were then identified. The first approach allows the identification of mucosa-associated domains from a set of previously known conserved regions, while the latter approach allows the discovery of new conserved regions associated to mucosal microorganisms. The identified conserved regions were then investigated further in order to generate hypotheses regarding their contribution to the survival of microbes in mucosal environments. The following sections describe these approaches in more detail.

6.2.1 Identification of mucosa-associated protein domains

To identify associations between protein domains and the habitat of microorganisms, a hypergeometric distribution test (see Section 2.9.1) was applied to all 8,423 InterPro (IPR) domains presented on 867 microorganisms' proteomes stored in the GenomePool database. The domain annotation results were produced by executing InterProScan as part of a high-throughput analysis workflow (see Chapter 3, Section 3.3.1). The habitat of organisms was annotated using information derived from the GOLD database [1] (downloaded 22$^{\text{nd}}$ October 2009) (See Section 5.2.1 and 5.3.1).

[1] http://www.genomesonline.org/, accessed 20th August 2010

Association analysis

The association analysis was used to determine the co-occurrence and the abundance of protein domains and the mucosal niches of microorganisms. To perform the association analysis, each taxa was assigned a binary classification. This classification either denotes the presence of that organism within a mucosal niche as a mutualist or pathogen, or alternatively indicates that there is no evidence of habitation in a mucosal environment. The classification was assigned to taxa according to the information available in the GOLD database. Three-hundred and forty-four taxa had isolation sources that were ambiguous. These taxa were removed from the analysis.

The hypergeometric test was used to assess the probability of finding a given protein domain in the test set in relation to the reference set. This statistical test was performed to assess two aspects of the mucosa-protein associations: the co-occurrence of the InterPro entries and mucosal microorganisms; and the abundance or expansion of the InterPro entries among mucosal microbes. The former aspect uncovers conserved protein sequence regions originating from lateral gene transfer (LGT) or gene loss events, whereas the latter case identifies functional regions arising from gene expansion or in combination with LGT events. The co-occurrence evaluation takes into account the presence or absence of a protein domain among mucosal and non-mucosal microorganisms. The abundance assessment takes into consideration the number of a given protein domain found across the two sets of microorganisms. More specifically for the co-occurrence assessment, hypergeometric probability distribution provides the probability (co-occurrence p-value) of observing the number of organisms within the test set (mucosa-associated microbes) with a given protein domain compared to the number habitat-classified organisms with that protein domain (reference set). To determine the abundance of a protein domain, the hypergeometric distribution provides the chance (abundance p-value) of observing the number of a given domain within the test set in comparison to the total number of that domain found in the reference set.

Moreover, the linear correlations between the protein domains and the ability of microorganisms to thrive on a mucosa-lined niche was also evaluated. The Pearson's product moment correlation coefficient (see Section 2.9.2) was employed to measure the correlation scores of each pair of a domain and the ability of microbes to thrive in mucosal environments.

Domain clustering

The abundance of IPR domains for each taxa were counted. Given a particular protein domain, these abundance values were normalised to have a mean of zero and a variance of one. A normalised value is a measure of relative abundance and depletion of a given domain across organisms. The data were clustered according to the profile of the protein domains by using Euclidean distance metric from the Cluster 3.0 application [de Hoon *et al.*, 2004]. Java Treeview [Saldanha, 2004] was used to visualise the results in a heatmap with the correspond dendrogram of variables.

Functional analysis and biological interpretation of protein domains

Domain descriptions from the InterPro database were used as a source of function information for each IPR domain of interest. If present, the GO term annotations of a protein domain were also employed for the identification of the three GO categories including: biological process, molecular function and cellular component. BiNGO [Maere *et al.*, 2005], a Cytoscape plug-in to assess GO term enrichment, was used to find statistically over-represented GO terms in a given set of InterPro protein domains. BiNGO was configured to use the hypergeometric test for the statistical test with an false discovery rate (FDR) correction for multiple testing. The significance level for inclusion was set to 0.05. The reference background annotations included the IPR domains of microorganisms in the data set from which the interested domains were identified as of interest.

6.2.2 Identification of mucosa-associated extracytoplasmic protein families

The purpose of the approach described in Section 6.2.1 was to identify mucosa-associated genetic elements from the set of previously known conserved protein regions represented in the InterPro database. However, it is anticipated that many more conserved regions have not yet been characterised, and are therefore not covered by any public protein domain databases. To address this issue, protein families of a set of known mucosa-thriving microbes were examined for their distribution among other mucosal microorganisms. The distribution of protein sequences among other mucosal organisms was evaluated by performing BLASTP searches of the sequences against all protein sequences in the RefSeq database. If the hit results from the all-vs-RefSeq BLASTP were significantly widely distributed across mucosa-dwelling organisms, it can be inferred that the function of the query protein might be associated with microbes' survival in mucosal environments. Applying this evaluation analysis on every protein sequence, the outcome therefore provides a list of candidate protein sequences that are potentially specific to mucosal microorganisms regarding the existence of their homologs across mucosa-associated taxa. Based on the construction of protein families and the known mucosa-associated protein candidate list, it is possible to identify groups of evolutionarily related extracytoplasmic proteins putatively shared across mucosal microbes. These homologous groups or protein families were then investigated further in order to determine their potential contribution to the adaptation of microbes to survive in mucosal environments. The approach allows the identification of groups of evolutionarily related proteins shared across mucosal microbes. From the protein families, it is then possible to reveal (where appropriate) the as yet undefined conserved regions that might be important for the survival of microbes in mucosal environments. The construction of the protein families focused on extracytoplasmic proteins of the 75 known mucosa-thriving microbes. The proteomes were from six different bacterial phyla: 5 Actinobacteria, 7 Bacteroidetes, 11 Chlamydiae, 15 Firmicutes, 1 Fusobacteria, 31 Proteobacteria, and 5 different protists (see Appendix D). In total 82,863 putative extracytoplasmic protein sequences out of 285,047 gene-coding protein sequences from the 75 organisms were included in the protein family construction.

Protein family construction

The set of 75 mucosa-adhering microbial extracytoplasmic proteomes of both mutualists and pathogens was clustered using OrthoMCL [Li *et al.*, 2003]. A pairwise all-against-all BLASTP analysis was performed using these proteomes to provide similarity scores between protein sequences. Protein sequences were then clustered into families based on their sequence similarity. The BLASTP pairwise results were retrieved from the in-house database storing output from the project Microbase workflow that executed the protein similarity searches (see Section 3.3.4).

OrthoMCL was employed to perform the clustering of homologous proteins with the inclusion criteria of a BLASTP e-value cut-off of $< 1 \times 10^{-5}$ and a percent identity cut-off of 50%. MCL was used with an inflation rate 1.5.

Identification of proteins overrepresented in mucosal microorganisms

To investigate the distribution of homologous sequences from other mucosal microorganisms not included in the 75 proteome data set, BLASTP was employed to search the 75 proteomes against all the sequences in RefSeq database. For each query sequence, BLAST hits with an e-value of $< 10^{-5}$ were investigated to determine whether their source organisms were mucosa-associated taxa. The hit source taxa were summarised in two different ways: the number of known mucosal organisms including both microbes and eukaryotic hosts (based on the information from the GOLD database); and the total number of taxa with a BLAST hit. If several proteins from the same taxon had positive BLAST hits, then these hits were counted as one. Based on these numbers, the p-value of the finding a hit sequence from the mucosal organisms by chance was calculated by using the hypergeometric distribution test. Hypergeometric mean values (see Section 2.9.1) were then used to infer the direction of the protein-mucosa association. The query sequences with a p-value of $< 1 \times 10^{-2}$ with positive associations to mucosal microbes were considered as the proteins specific to mucosal microorganisms and therefore potentially important for the mucosa-microbe interactions.

Functional annotation of protein clusters

BLASTP was used to examine the functional differences across the generated protein families. The similarity searches were performed on sequences in each family against a set of proteins of known functional annotations from Clusters of Orthologous Groups (COG) [Tatusov *et al.*, 2000] for the prokaryotic proteins, and the eukaryotic Orthologous Groups (KOG) [Tatusov *et al.*, 2003] for the proteins of microbial eukaryotes. The best BLAST hits and with an e-value threshold of less than 1×10^{-10} for all sequences in a cluster were used to assign the COG or KOG family to a protein cluster.

6.3 Results

6.3.1 Comparative genomics to reveal niche-specific protein domains

The distribution of each of 8,423 IPR domains on all 867 microorganisms was investigated to examine their conservation and abundance among microorganisms from different niches. The hypergeometric test was used to identify significant associations between protein domains and particular habitats. This method revealed several sets of IPR domains to be significantly associated with different ecological niches. Among the 8,000 IPR domains, 231 were determined to be significantly associated with organisms thriving in a mucosal environment (co-occurrence p-value $< 10^{-4}$). The set of 231 domains were conserved mainly among microorganisms isolated from mucosal environments (see Figure 6.1). The specification of the 231 mucosa-associated domains were found to be spread across different type of mucosa niches. Moreover, different types of symbiotic relationships (i.e., pathogenic, mutual) between the microbe and host appear to have different sets of conserved domains. For example, inhibitors of vertebrate lysozyme (IPR014453) are found exclusively among pathogenic Proteobacteria.

Some domains that are significantly overrepresented in soil-living microorganisms are also predominant in gastrointestinal tract pathogens as several soil-based organisms are pathogenic to mammal hosts once coming into contact with mucosa surfaces. For example, *Bacillus cereus* is regarded

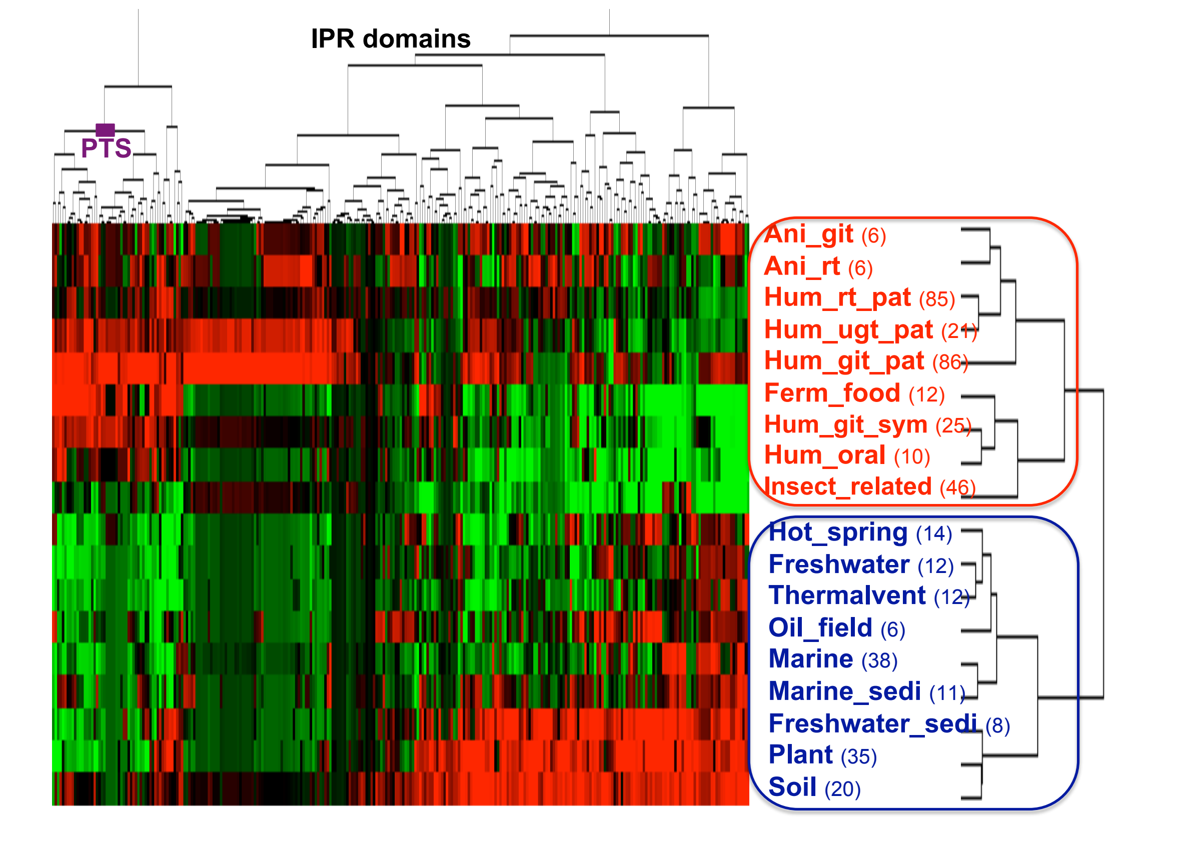


Figure 6.1: Distribution of InterPro domains from microorganisms in different habitats. The heatmap dendrogram shows the abundance of the domains in microorganisms from different habitats, calculated using a (centered-mean) normalisation of the percent coverage of the selected protein domains across the different habitats of the taxa. Different symbiosis relationships (pathogenic or mutual) between of microbes and host-associated habitats were also indicated where possible. The IPR domains significantly associated with taxa surviving on mucosal surfaces were selected, as well as the contrasting set of domains that are strongly associated with soil-dwelling microbes. For a given domain, red shows a larger proportion of taxa having that domain and living in a given habitat, whereas green shows smaller proportion of taxa that have that domain. The hierarchical clustering was performed using the complete linkage method and Euclidean distance based similarity. Numbers in brackets indicate the number of organisms living in a particular habitat that were analysed. PTS=sugar phosphotransferase system, Ani=animal, Hum=human, git=gastrointestinal tract, rt=respiratory tract, ugt=urogenital tract, pat= pathogen, sym=symbiont, sedi=sediment. Of particular interest are the domains shown to the left of the diagram. These domains are present mainly in the organisms that are associated with mucosal surfaces. In contrast, the domains to the right of the diagram are present more in microorganisms that are isolated from other environments.

as a soil-dwelling microbe, but several strains of *B. cereus* are occasionally found as the cause of diarrhoea in humans [Arnesen *et al.*, 2008]. The overall functions of the soil-associated protein domains were investigated further in relation to their GO annotations. Soil-specific protein domains are involved in the biosynthesis of a coenzyme (pantothenate), nitrogen compounds and the histidine family amino acids (see Figure 6.2). The significant molecular functions of these protein domains plays a role in electron carrier activity, histidinol dehydrogenase activity, copper ion binding and iron-sulphur cluster binding. A detailed analysis of mucosa-specific protein domains is described in the next section.

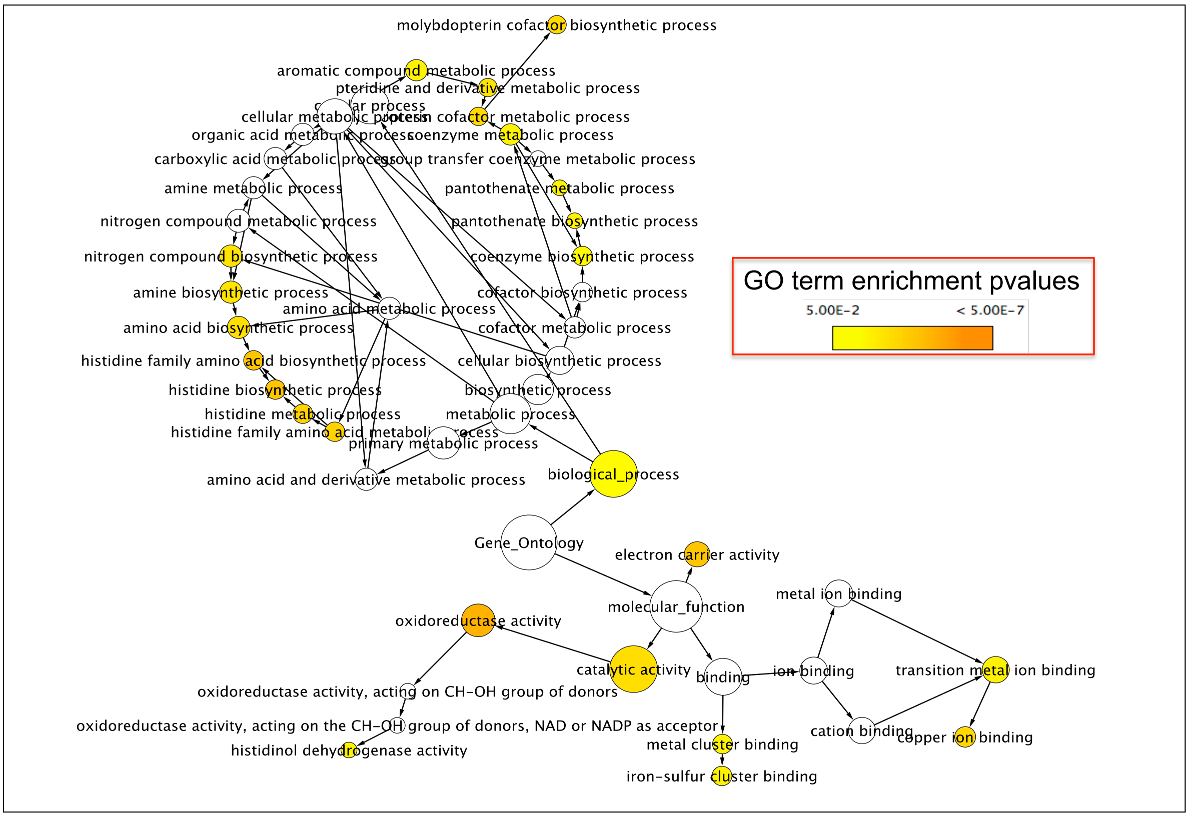

molybdopterin cofactor biosynthetic process
aromatic compound metabolic process
pteridine and derivative metabolic process
cellular process
cellular metabolic process
pterin cofactor metabolic process
organic acid metabolic process
coenzyme metabolic process
carboxylic acid metabolic process
group transfer coenzyme metabolic process
amine metabolic process
pantothenate metabolic process
nitrogen compound metabolic process
pantothenate biosynthetic process
nitrogen compound biosynthetic process
coenzyme biosynthetic process
amine biosynthetic process
cofactor biosynthetic process
amino acid metabolic process
amino acid biosynthetic process
cofactor metabolic process
histidine family amino acid biosynthetic process
cellular biosynthetic process
histidine biosynthetic process
biosynthetic process
histidine metabolic process
metabolic process
histidine family amino acid metabolic process
primary metabolic process
amino acid and derivative metabolic process
biological_process
Gene_Ontology
electron carrier activity
metal ion binding
molecular_function
oxidoreductase activity
catalytic activity
binding
ion binding
transition metal ion binding
oxidoreductase activity, acting on CH-OH group of donors
oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor
cation binding
copper ion binding
histidinol dehydrogenase activity
metal cluster binding
iron-sulfur cluster binding
GO term enrichment pvalues
5.00E-2
< 5.00E-7

Figure 6.2: GO terms overrepresented among soil-associated protein domains. A set of 161 InterPro protein domains were found to significantly coexist with microorganisms isolated from soil (p-value $< 1 \times 10^{-5}$). White nodes are GO terms with no significant enrichment, but are included because they have a significant child term. The size of each node is proportional to the number of nodes in the data set with a given GO term.

6.3.2 Protein domains overrepresented in mucosa-thriving microbes

This section describes the detailed analysis of mucosa-specific protein domains. All mucosa-thriving and non-mucosa-dwelling microbes from across the three domains of cellular life were used. Where multiple strains of a particular species exist, only one of the most well-known strain was selected. In the case where different strains are isolated from different sources, one well-known strain of each isolation source was included. These processes was performed in order to reduce noise and bias that may occur by several copies of nearly identical genome sequences. As a result, after the removal of the redundant proteomes, 463 microorganism proteomes remained, of which 122 were annotated as mucosal-thriving microbes and 341 were marked as non-mucosa associated taxa (see Figure 6.3). In total, there are 8,243 InterPro entries annotated on the selected 463 microorganism proteomes included in this analysis.

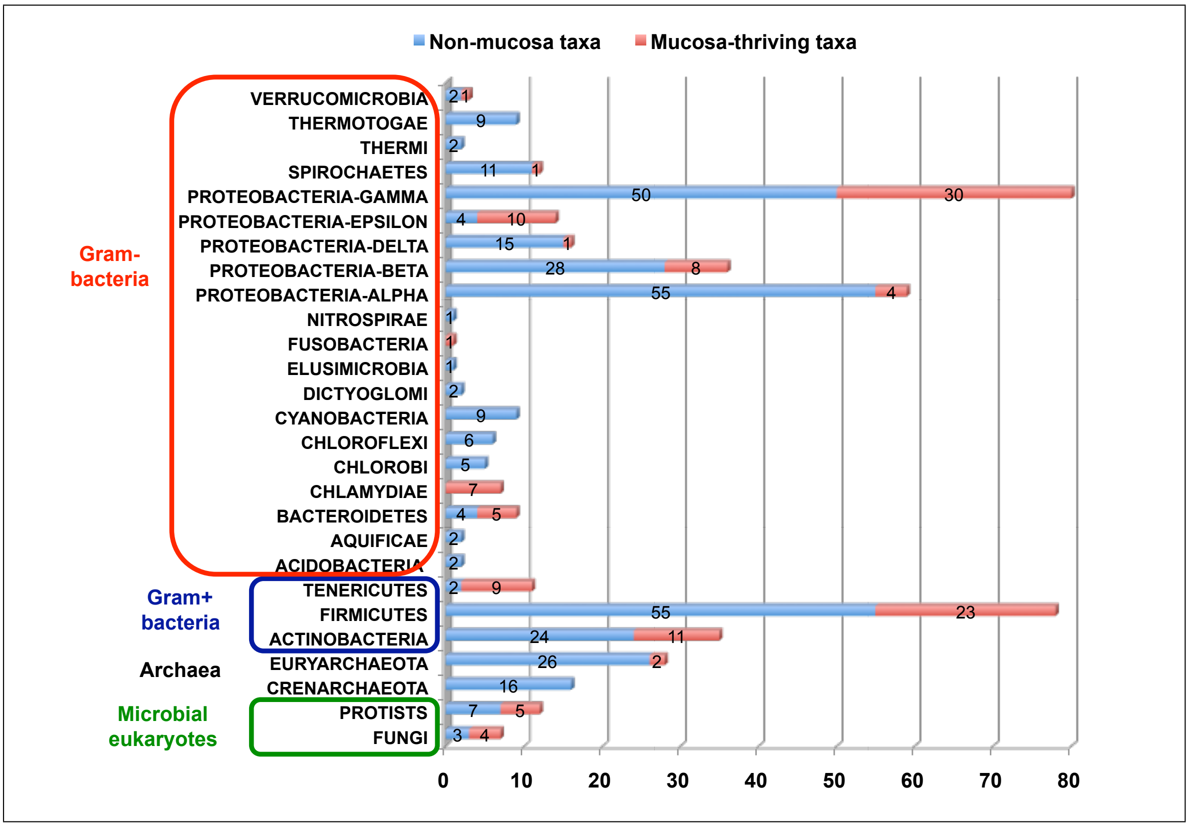


Figure 6.3: Distribution of 463 microbial mucosal and non-mucosal taxa across the NCBI taxonomic classification. Red shows the number of mucosa-thriving taxa, whereas blue indicates the number of non-mucosa taxa. The proteomes of both mucosa and non-mucosa microorganisms included in the analysis are distributed across taxonomic tree. Notably, most of taxa are Proteobacteria and Firmicutes.

The association and functional analyses of protein domains overrepresented in mucosa-thriving microorganisms

This section describes the approach that was used to investigate the overall molecular functions common across the microorganisms capable of thriving on mucosa-lined niches. The association analysis in this section was performed on a wide set of microorganism proteomes including the data set from bacteria, archaea and microbial eukaryotes. Localisation of protein sequences were not taken into account in this analysis. A significance test was applied and an association score was computed for each InterPro entry to determine if the entry was significantly present in the mucosal taxa compared to the occurrence of the entry among the non-mucosa taxa.

Figure 6.4 shows the distribution of all 8,243 InterPro domains across mucosal and non-mucosa taxa with respect to the percent coverage that a given domain occurs in both sets of taxa. At a cutoff (uncorrected) p-value of $< 1 \times 10^{-4}$ (see Figure 6.5), 231 InterPro domain entries appeared to be statistically associated with microbes thriving in a mucosal environment. The direction of the association was determined by a Hypergeometric mean value (see Section 2.9.1) and Pearson's correlation coefficient (see Section 2.9.2). A more stringent p-value was also considered in order to remove false negatives that may occur from the multiple hypothesis test. Using a cut-off p-value 1×10^{-5}), 119 out of 231 InterPro entries had passed this cut-off value. To investigate the biological meaning of these mucosa-associated protein domains, Gene Ontology (GO) terms enrichment assessment was performed to pinpoint GO terms that were overrepresented. Interestingly, the result set obtained by using both cut-off p-values provides the same overview of GO term enrichment with a slight difference in their p-values yielded from the GO enrichment analysis. Summary results from the GO terms enrichment analysis are shown in Figures 6.6, 6.7 and 6.8 for cellular component, biological process and molecular function, respectively.

The results show that domains that are overrepresented among mucosa-thriving microbes are possessed by cell membrane and cell wall proteins. Those protein sequences appear to be involved mainly in carbohydrate and amine transport activities, especially sugar transport via the phosphotransferase system (PTS)[Postma *et al.*, 1993]. They are also generally involved in cell communication, signal transductions establishment of localisation, and biological regulation [Houot *et al.*, 2010] [Gosset, 2005] [Vadeboncoeur and Pelletier, 1997]. The PTS is one of the main carbohydrate transport systems in bacteria[Postma *et al.*, 1993]. Interestingly, when using a less stringent p-value cutoff ($< 1 \times 10^{-2}$) for the inclusion criteria of protein domains for the GO terms enrichment analysis, terms under cellular metabolic processes such as carbohydrate and alcohol metabolic processes also appeared to be overrepresented among the mucosal taxa data set. This interpretation of the results obtained corresponds with the recent metagenomics analysis of the human distal gut microbiome which identified the biodegradation of complex sugars and glycans as an important function for life of gut bacteria [Gill *et al.*, 2006]. From our analysis, the results suggest that the functions involved with the carbohydrate transport may also be important among the non-gut mucosa-thriving microbes as well, since the PTS-related protein domains are distributed across mucosa-associated microorganisms both symbionts and pathogens (see Figure 6.1). The PTS-related domains were found in microorganisms known to be able to thrive in human oral, urogenital and respiratory tracts (see Figure 6.17). The results suggest that the PTS is an important system that enables bacteria to respond to the availability of carbohydrate substrates by using them as preferred carbon sources. The PTS transports sugars aiding to the microorganisms' survival in carbohydrate-rich environments such as mucosa-coated surfaces [Houot *et al.*, 2010] [Gosset, 2005] [Vadeboncoeur and Pelletier, 1997].

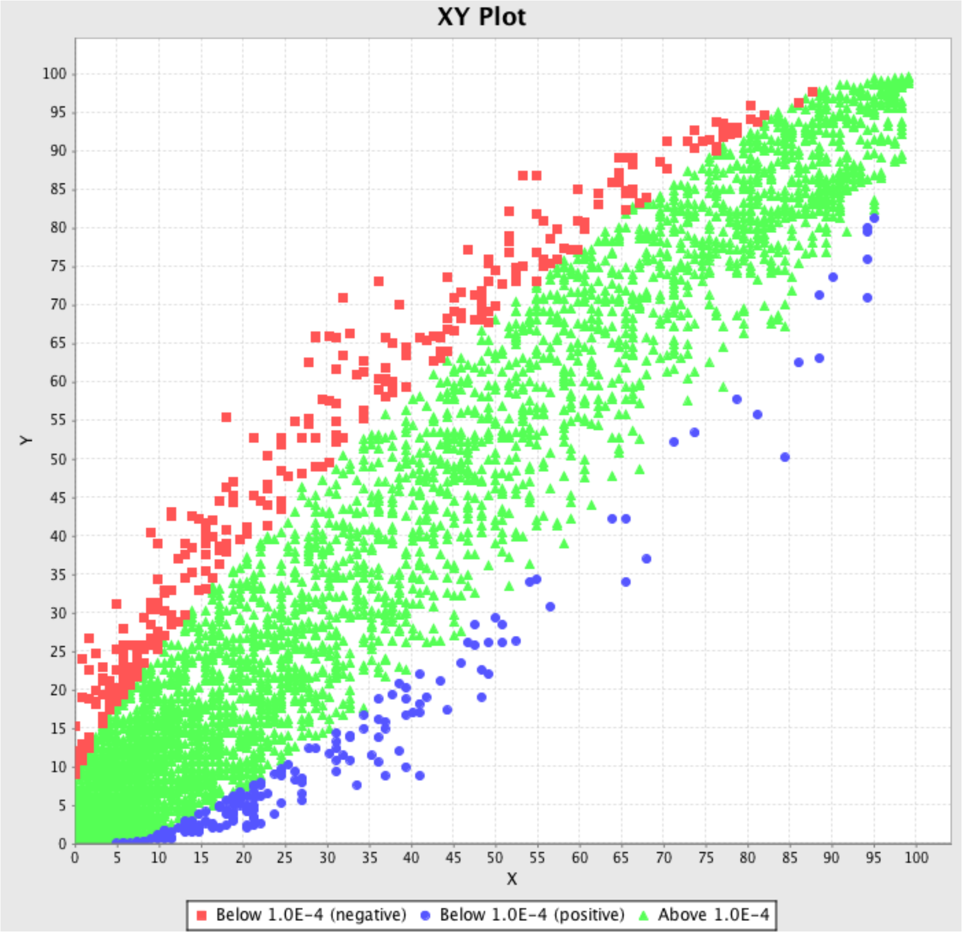


Figure 6.4: Distribution of 8,243 InterPro domain entries among the set of mucosa-thriving taxa and non mucosa-associated taxa. Plots represent InterPro protein domains. The X-axis and Y-axis are percentages of known mucosa-thriving and non mucosa-associated microorganisms that have a given protein domain, respectively. The colours of the plots show the level of significance p-values obtained from the association analysis. The colour representing the level of p-value is shown in the text box below the plot.

In addition to the protein domains significantly associated with mucosa-thriving microbes, the domains that were significantly negative or were underrepresented in the mucosal taxa were also investigated. The domains deprived in mucosal microbes compared to the set of non-mucosal microbes are mainly proteins associated with plastids and chloroplasts. These proteins are involved in activities involving inorganic compounds such as metal, copper-ion and vitamin binding activities as well as catalytic activities including oxidoreductase, ligase and carboxy-lyase activities. The overall biological process of these non-mucosa associated protein domains are vitamin, cofactor and heterocycle metabolic processes as well as metabolic processes for carbon utilisation, response to external stimulus and oxidation reduction such as electron transport. One explanation for this variety may be that the non-mucosal taxa set were free living microbes that survive in soil, plants, marine environments,

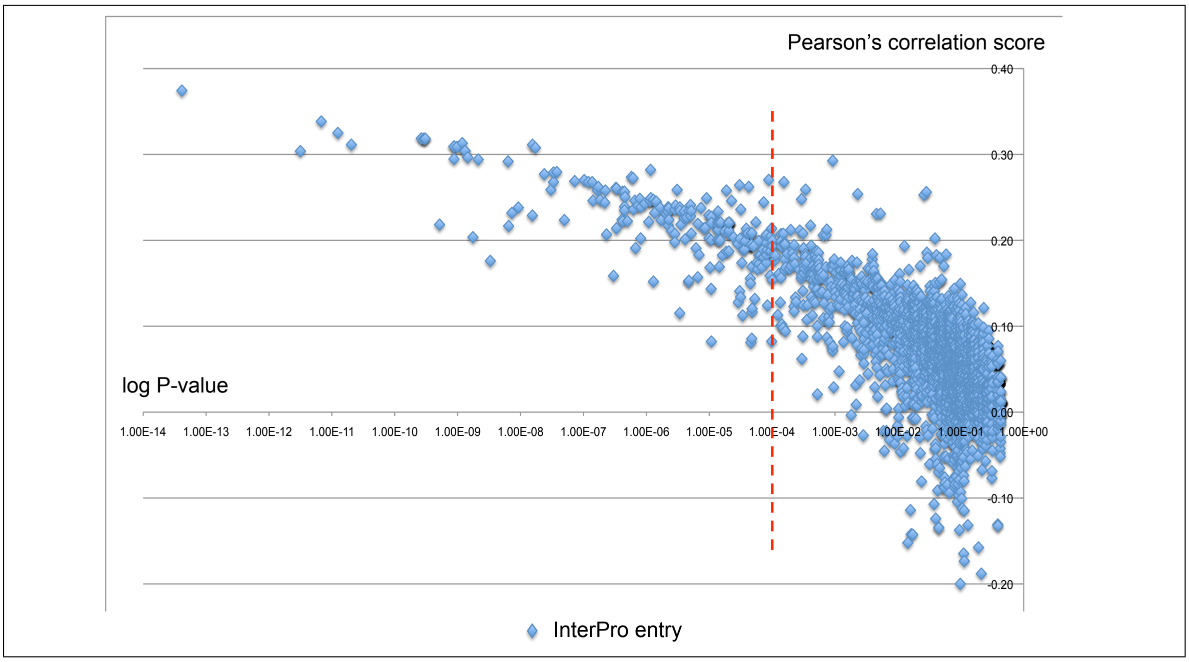


Figure 6.5: Plots of protein domains showing a positive association with mucosal microorganisms. The direction of the association was determined by hypergeometric mean values. The red dashed line indicates the cutoff p-value used as an inclusion criteria to declare protein domains as significant associations.

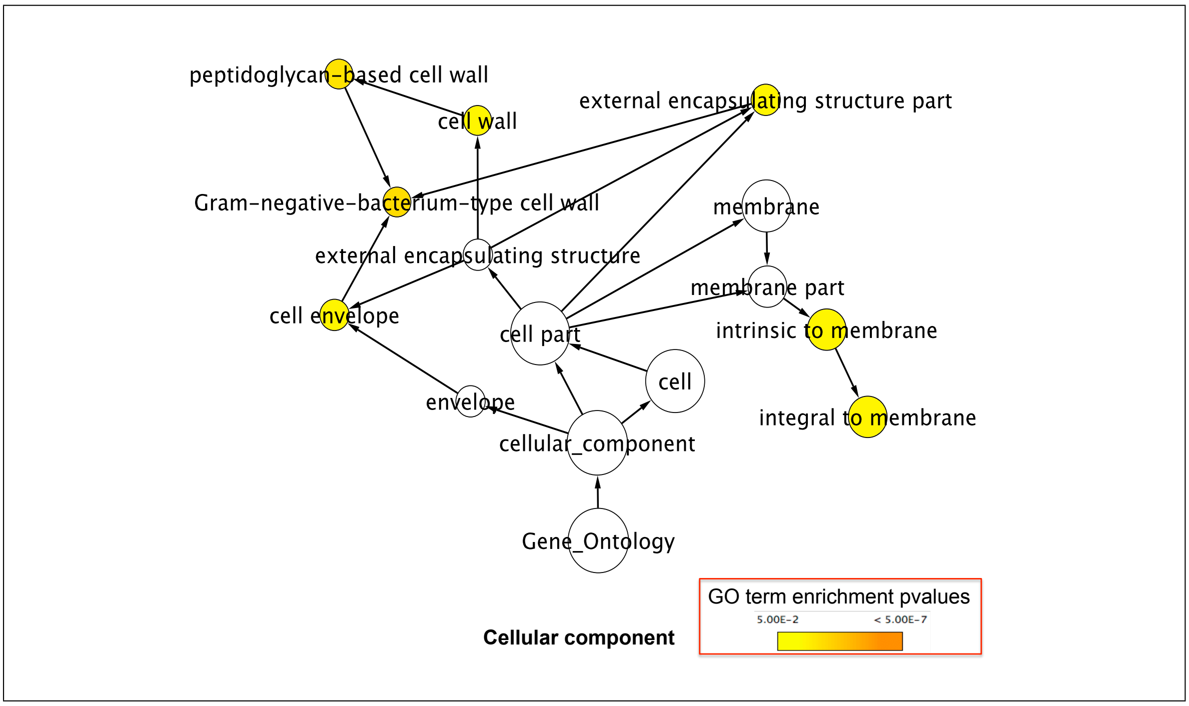


Figure 6.6: Cellular component GO terms overrepresented among InterPro protein domains that were statistically significantly associated with mucosal taxa. These GO terms are enriched among the set of IPR domains marked in blue in Figure 6.4. White nodes are terms with no significant enrichment, but are included because they have a significant child term. The size of each node is proportional to the number of nodes in the data set with a given GO term.

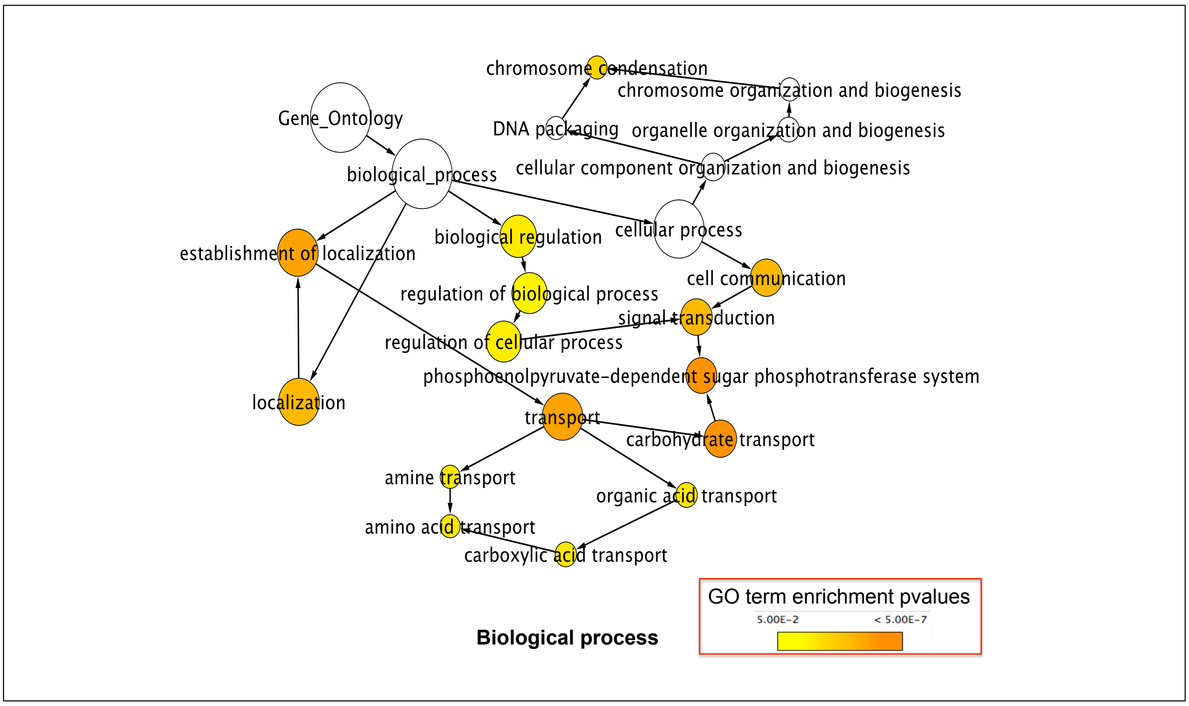

chromosome condensation
chromosome organization and biogenesis
Gene_Ontology
DNA packaging
organelle organization and biogenesis
biological_process
cellular component organization and biogenesis
biological regulation
cellular process
establishment of localization
regulation of biological process
cell communication
signal transduction
regulation of cellular process
phosphoenolpyruvate-dependent sugar phosphotransferase system
localization
transport
carbohydrate transport
amine transport
organic acid transport
amino acid transport
carboxylic acid transport
GO term enrichment pvalues
5.00E-2
< 5.00E-7
Biological process

Figure 6.7: Biological process GO terms overrepresented among InterPro protein domains that were statistically significantly associated with mucosal taxa. These GO terms are enriched within the set of IPR domains marked in blue in Figure 6.4.

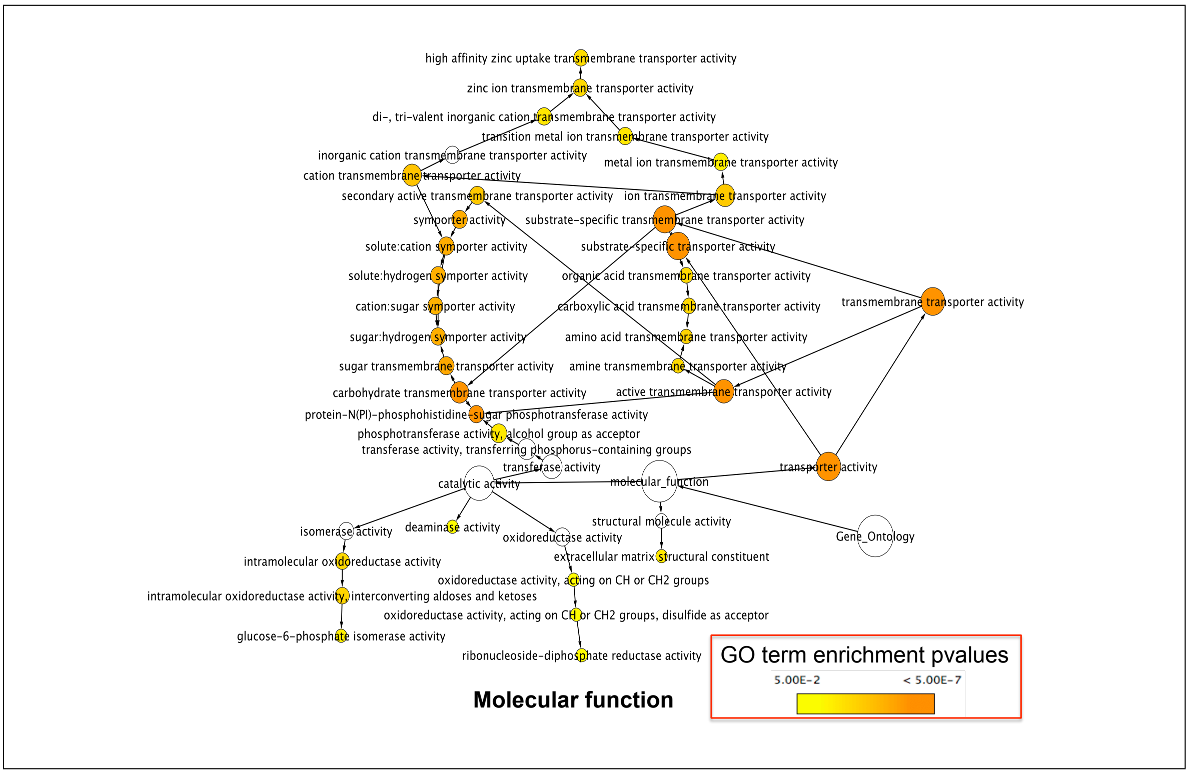

high affinity zinc uptake transmembrane transporter activity
zinc ion transmembrane transporter activity
di-, tri-valent inorganic cation transmembrane transporter activity
transition metal ion transmembrane transporter activity
inorganic cation transmembrane transporter activity
metal ion transmembrane transporter activity
cation transmembrane transporter activity
secondary active transmembrane transporter activity
ion transmembrane transporter activity
symporter activity
substrate-specific transmembrane transporter activity
solute:cation symporter activity
substrate-specific transporter activity
solute:hydrogen symporter activity
organic acid transmembrane transporter activity
cation:sugar symporter activity
carboxylic acid transmembrane transporter activity
transmembrane transporter activity
sugar:hydrogen symporter activity
amino acid transmembrane transporter activity
sugar transmembrane transporter activity
amine transmembrane transporter activity
carbohydrate transmembrane transporter activity
active transmembrane transporter activity
protein-N(PI)-phosphohistidine-sugar phosphotransferase activity
phosphotransferase activity, alcohol group as acceptor
transferase activity, transferring phosphorus-containing groups
transferase activity
transporter activity
catalytic activity
molecular_function
isomerase activity
deaminase activity
oxidoreductase activity
structural molecule activity
Gene_Ontology
intramolecular oxidoreductase activity
extracellular matrix structural constituent
oxidoreductase activity, acting on CH or CH2 groups
intramolecular oxidoreductase activity, interconverting aldoses and ketoses
oxidoreductase activity, acting on CH or CH2 groups, disulfide as acceptor
glucose-6-phosphate isomerase activity
ribonucleoside-diphosphate reductase activity
GO term enrichment pvalues
5.00E-2
< 5.00E-7
Molecular function

Figure 6.8: Molecular function GO terms overrepresented among InterPro protein domains that were statistically significantly associated with mucosal taxa. These GO terms are enriched within the set of IPR domains marked in blue in Figure 6.4.

deep seas and hot springs. These microorganisms therefore acquire energy from various sources depending on their surrounding environment. However, carbohydrate and amino acid transport and metabolic processes were not overrepresented among these free-living microbes.

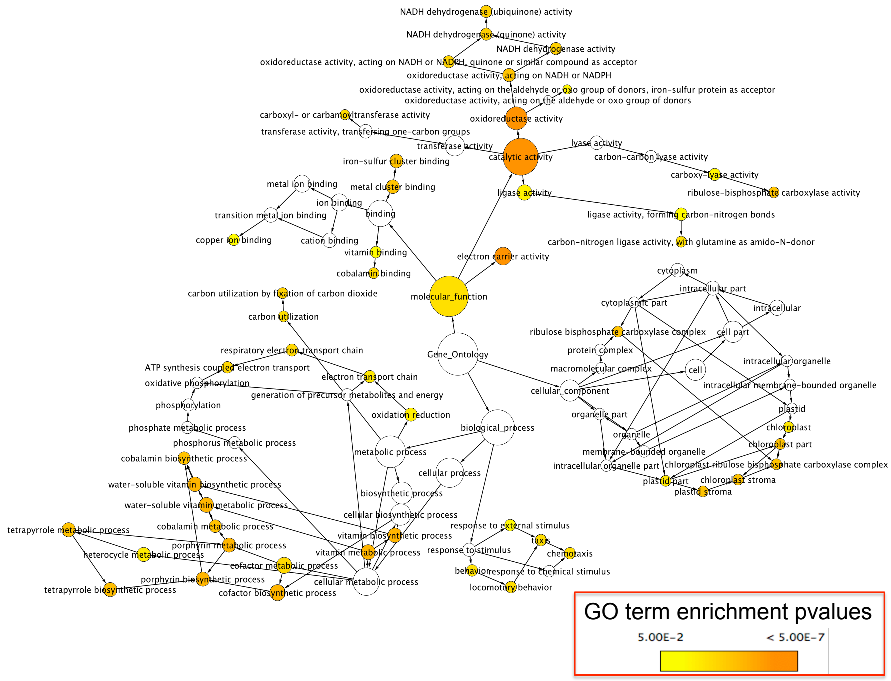

NADH dehydrogenase (ubiquinone) activity
NADH dehydrogenase (quinone) activity
NADH dehydrogenase activity
oxidoreductase activity, acting on NADH or NADPH, quinone or similar compound as acceptor
oxidoreductase activity, acting on NADH or NADPH
oxidoreductase activity, acting on the aldehyde or oxo group of donors, iron-sulfur protein as acceptor
oxidoreductase activity, acting on the aldehyde or oxo group of donors
carboxyl- or carbamoyltransferase activity
oxidoreductase activity
transferase activity, transferring one-carbon groups
transferase activity
lyase activity
catalytic activity
carbon-carbon lyase activity
iron-sulfur cluster binding
carboxy-lyase activity
metal ion binding
metal cluster binding
ligase activity
ribulose-bisphosphate carboxylase activity
ion binding
transition metal ion binding
binding
ligase activity, forming carbon-nitrogen bonds
copper ion binding
cation binding
carbon-nitrogen ligase activity, with glutamine as amido-N-donor
vitamin binding
electron carrier activity
cobalamin binding
cytoplasm
intracellular part
carbon utilization by fixation of carbon dioxide
molecular_function
cytoplasmic part
intracellular
carbon utilization
ribulose bisphosphate carboxylase complex
cell part
respiratory electron transport chain
Gene_Ontology
protein complex
ATP synthesis coupled electron transport
intracellular organelle
macromolecular complex
cell
electron transport chain
oxidative phosphorylation
cellular_component
intracellular membrane-bounded organelle
generation of precursor metabolites and energy
phosphorylation
oxidation reduction
organelle part
plastid
phosphate metabolic process
biological_process
chloroplast
organelle
phosphorus metabolic process
chloroplast part
metabolic process
membrane-bounded organelle
cobalamin biosynthetic process
intracellular organelle part
chloroplast ribulose bisphosphate carboxylase complex
cellular process
plastid part
chloroplast stroma
water-soluble vitamin biosynthetic process
biosynthetic process
plastid stroma
water-soluble vitamin metabolic process
cellular biosynthetic process
response to external stimulus
tetrapyrrole metabolic process
cobalamin metabolic process
vitamin biosynthetic process
taxis
porphyrin metabolic process
heterocycle metabolic process
vitamin metabolic process
response to stimulus
chemotaxis
cofactor metabolic process
behavior
response to chemical stimulus
porphyrin biosynthetic process
cellular metabolic process
tetrapyrrole biosynthetic process
cofactor biosynthetic process
locomotory behavior
GO term enrichment pvalues
5.00E-2
< 5.00E-7

Figure 6.9: GO terms underrepresented among InterPro protein domains that were statistically significantly associated with mucosal taxa compared to the non-mucosa proteome data set. This GO graph represents the GO term enrichment among the set of InterPro entries plotted in red in the Figure 6.4 (i.e., those with a p-value < 1 $\times 10^{-4}$ and has a negative association with mucosal organism data set).

Detailed functional analysis of the identified mucosa-associated protein domains

Investigations into the 231 mucosa-associated domains in order to identify their major functions and involvement in mucosa-microbe interactions were then carried out. Interestingly, these 231 domains not only co-occur with microorganisms that can thrive on host mucosa surfaces, but also they appear to be abundant among the mucosal microorganisms (abundance p-value ranging from 10^{-3} to 10^{-97}) (see Table 6.1, 6.2, Appendix G). All of these domains have patchy distribution among a specific taxonomic group of the annotated mucosal microorganisms (see Figure 6.10), suggesting specific groups of taxa have exclusive sets of protein domains.

Most of the identified domains are specific to bacteria. For example, PTS-related domains were distributed across Gram-positive bacteria and some Gram-negative bacterial phyla but are not found in microbial eukaryotes or archaea. Inhibitors of vertebrate lysozyme (IPR014453) are found exclu-

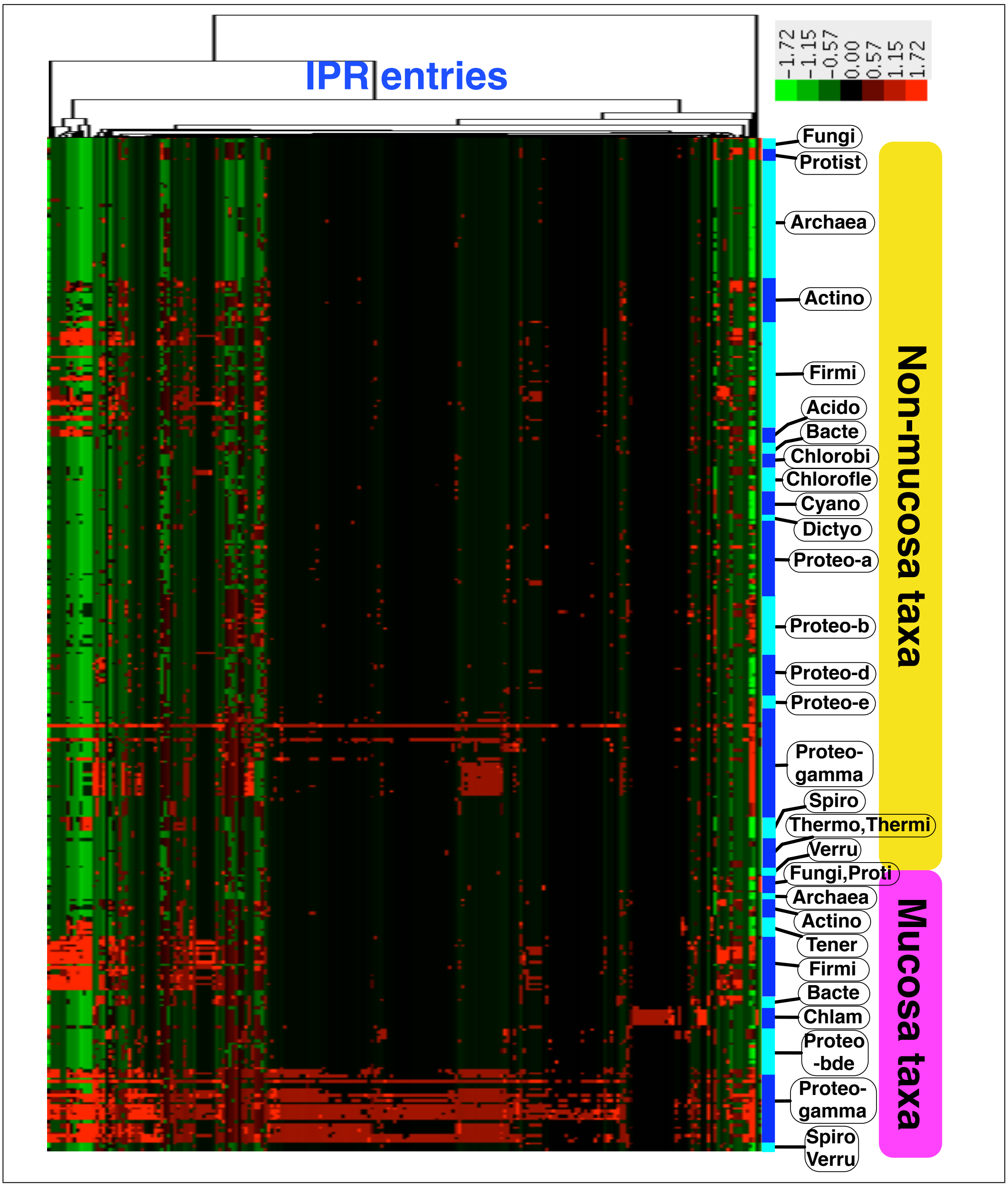


Figure 6.10: Distribution of mucosa-associated protein domains across microbial taxa. The heatmap dendrogram shows a normalised percentage coverage of each domain across different taxa. Each column represents an InterPro protein domain. Taxa (rows) were split into two different groups (Mucosal and non-mucosal taxa) in respect to their ability to thrive on mucosal surfaces. The colour coding indicates enrichment (red) and depletion (green) of a domain in a given taxa in relation to other taxa having that domain. Black shows an absence of a domain. The heatmap shows that these domains are overrepresented among taxa known to thrive in mucosal environments, particularly, mucosa-associated Proteobacteria-Gamma and Firmicutes. The domains are shown in blue circle in Figure 6.4.

sively among Proteobacteria (alpha, beta and gamma) whose members are often known as mucosa-associated pathogens. Among the 231 strongly mucosa-associated IPR domains, 19 entries were found to be shared across members of the three domains of life (archaea, bacteria and eukaryotes). Several were involved in DNA and RNA metabolic processes. The Glycosyl hydrolase family 32 (IPR013189, IPR013148, IPR001362) performs glycolysis activity by hydrolysing O-glycosyl compounds. This family appeared across eukaryotes (Kinetoplastida, Parabasalidea, Ascomycota, Basidiomycota), and bacteria (Halobacteria, Acidobacteria, Actinobacteria, Bacteroidetes, Dictyoglomi, Proteobacteria, Planctomycetes, Fusobacteria, Chloroflexi, Tenericutes, Thermotogae, Spirochaetes, and Verrucomicrobia). Most of the microbes carrying the Glycosyl hydrolase family 32 domain are able to thrive on various host mucosa surfaces. For example, *Brachyspira murdochii DSM 1256* is considered as a swine intestinal commensal; *Fusobacterium nucleatum subsp. polymorphum* ATCC 10953 is associated with human periodontal disease; *Corynebacterium urealyticum* is known to cause human urinary tract infection; *Vibrio cholerae* is a gastrointestinal pathogen; and *Trichomonas vaginalis* G3 is known as a sexually transmitted parasite that is able to colonise human urogenital tract mucosa [Hirt *et al.*, 2002]. Other interesting domains that are significantly overrepresented in mucosal microorganisms and also distributed across the three domains of life were Peptidase C69, dipeptidase A (IPR005322) and bacterial adhesion (IPR008966). More details of these domains are described in Section 6.4.1.

Out of 231 IPR domains, 64 (27.7%) domains that are strongly associated with mucosa-dwelling microbes were of unknown function (see Appendix G), whereas 84 (36.4%) IPR domains were characterised but as yet not annotated with GO terms (see Table 6.3). The remaining 83 mucosa-associated IPR domains were annotated with GO terms, of which 53 domains were annotated with the GO biological process terms that were statistically overrepresented within mucosa-thriving microorganisms (with co-occurrence p-value $< 1 \times 10^{-2}$) (see Table 6.1 and 6.2).

Using the approach described in this section, several known proteins domains assisting the survival of microbes on mucosal environment were identified. These domains are involved with a number of processes such as sensing carbohydrate-enriched environment, carbohydrate metabolic processes, sugar translocation, adhesion, responding to acidity and other stress conditions, proteolysis, host anti-bacterial inhibition, and pathogenesis. Opacity-associated protein A (OapA; IPR013731, IPR007340) was a characterised domain that was identified by the approach described above (see Section 6.2.1) as a mucosa-associated protein domain (co-occurrence p-value 3.3×10^{-5}, abundance p-value 8.9×10^{-8}). OapA is known to contribute to efficient colonisation of *Haemophilus influenzae* to the nasopharyngeal mucosa. The choloylglycine hydrolase domain (IPR003199) is a known gut-specific domain that was also identified by the approach used in this study as a mucosa-associated domain with significant co-occurrence and abundance p-values (co-occurrence p-value 1.9 $\times 10^{-4}$, abundance p-value 6.9 $\times 10^{-6}$).

The domains involved in phosphoenolpyruvate-dependent PTS appear to play a major role in the signal transduction and carbohydrate transport among mucosa-thriving bacteria. Not surprisingly, the PTS regulation domain, PRD, was also observed as a highly abundant mucosa-associated domain (co-occurrence p-value = 3 $\times 10^{-7}$, abundance p-value = 2.8 $\times 10^{-23}$). The PRDs, common in Gram-positive bacteria, are found in both bacterial transcriptional anititerminators and activators which are modulated by phospholyration [Stülke *et al.*, 1998]. In the presence of PTS substrates (carbohydrates), the PRD-containing regulators activate the expression of operons or genes involved in carbohydrate transport by the PTS. While lacking an inducer, the PRD regulator stimulates the generation of PTS substrates. The PRD regulator has been found to be inhibited in the presence of a rapid metabolisable carbon source [Stülke *et al.*, 1998].

Table 6.1: **A list of mucosa-associated IPR domains annotated with overrepresented GO biological process.** These GO biological processes terms were overrepresented in the GO term enrichment analysis of IPR domains, with a co-occurrence p-value < 1 $\times 10^{-2}$. The entries were categorised based on their corresponding GO terms.

GO Biological process	IPR entry	co-occurrence p-value	abundance p-value	correlation score
Phosphoenolpyruvate-dependent sugar phosphotransferase system (PTS) (3.36 $\times 10^{-14}$)				
PTS, lactose/cellobiose-specific IIB subunit	IPR003501	9.21 $\times 10^{-9}$	1.45 $\times 10^{-35}$	0.24
PTS, mannose/fructose/sorbose family IID component	IPR004704	6.14 $\times 10^{-7}$	1.05 $\times 10^{-25}$	0.27
PTS, sorbose-specific IIC subunit	IPR004700	5.80 $\times 10^{-7}$	3.26 $\times 10^{-26}$	0.27
PTS, glucitol/sorbitol-specific IIA component	IPR004716	1.67 $\times 10^{-7}$	1.67 $\times 10^{-9}$	0.26
PTS, sorbose subfamily IIB component	IPR004720	1.03 $\times 10^{-7}$	3.20 $\times 10^{-25}$	0.27
PTS, galactitol-specific IIC component	IPR004703	8.25 $\times 10^{-6}$	5.75 $\times 10^{-9}$	0.21
PTS, sugar-specific permease EIIA 1 domain	IPR001127	7.38 $\times 10^{-6}$	1.69 $\times 10^{-23}$	0.24
PTS, EIIB component, type 3	IPR013012	4.74 $\times 10^{-6}$	5.34 $\times 10^{-11}$	0.15
PTS, lactose/cellobiose-specific IIA subunit	IPR003188	4.66 $\times 10^{-6}$	1.10 $\times 10^{-9}$	0.15
PTS, EIIB	IPR001996	3.08 $\times 10^{-6}$	1.60 $\times 10^{-40}$	0.26
Sorbitol phosphotransferase enzyme II, N-terminal	IPR011618	2.23 $\times 10^{-6}$	3.49 $\times 10^{-7}$	0.23
Sorbitol phosphotransferase enzyme II, C-terminal	IPR011638	2.23 $\times 10^{-6}$	2.21 $\times 10^{-7}$	0.22
PTS, enzyme II sorbitol-specific factor	IPR004699	2.23 $\times 10^{-6}$	3.49 $\times 10^{-7}$	0.23
PTS, EIIB component, type 2	IPR013011	3.03 $\times 10^{-5}$	9.31 $\times 10^{-47}$	0.26
Other carbohydrate transport (1.40 $\times 10^{-13}$)				
Maltose operon periplasmic	IPR010794	2.28 $\times 10^{-6}$	2.10 $\times 10^{-6}$	0.22
ABC transporter, maltose/maltodextrin import, MalK	IPR015855	1.14 $\times 10^{-6}$	1.25 $\times 10^{-6}$	0.25
amino acid and carboxylic acid transport (4.26 $\times 10^{-3}$)				
Branched-chain amino acid transport system II carrier protein	IPR004685	5.16 $\times 10^{-10}$	9.71 $\times 10^{-12}$	0.22
ABC transporter, methionine import, ATP-binding protein, MetN, C-terminal	IPR017908	1.75 $\times 10^{-9}$	8.91 $\times 10^{-7}$	0.20
Anaerobic c4-dicarboxylate membrane transporter	IPR004668	1.45 $\times 10^{-9}$	6.39 $\times 10^{-16}$	0.30
Sodium/glutamate symporter	IPR004445	6.73 $\times 10^{-7}$	1.28 $\times 10^{-6}$	0.19
chromosome condensation (6.67 $\times 10^{-4}$)				
Prokaryotic chromosome segregation and condensation protein MukE	IPR007385	8.75 $\times 10^{-10}$	1.81 $\times 10^{-9}$	0.31
Prokaryotic chromosome segregation and condensation protein MukB, N-terminal	IPR007406	2.90 $\times 10^{-10}$	6.24 $\times 10^{-10}$	0.32
Histone H1-like nucleoprotein HC2	IPR009970	2.92 $\times 10^{-6}$	2.62 $\times 10^{-6}$	0.24
Glucose metabolic process				
Phosphoglucose isomerase (PGI)	IPR001672	5.30 $\times 10^{-5}$	7.24 $\times 10^{-3}$	0.13
Phosphoglucose isomerase, conserved site	IPR018189	3.41 $\times 10^{-5}$	9.13 $\times 10^{-3}$	0.11
Pyruvate formate-lyase, PFL	IPR004184	2.53 $\times 10^{-6}$	7.65 $\times 10^{-12}$	0.21
Other monosaccharide metabolic process				
L-fucose isomerase, C-terminal	IPR004216	3.63 $\times 10^{-5}$	6.77 $\times 10^{-5}$	0.19
Other carbohydrate metabolic process				
Mannose-6-phosphate isomerase, type I	IPR001250	5.00 $\times 10^{-7}$	1.55 $\times 10^{-6}$	0.22
Putative N-acetylmannosamine-6-phosphate epimerase	IPR007260	1.18 $\times 10^{-9}$	3.26 $\times 10^{-10}$	0.31
Glucosamine/galactosamine-6-phosphate isomerase	IPR006148	2.45 $\times 10^{-5}$	3.05 $\times 10^{-7}$	0.20
Glycoside hydrolase, family 32	IPR001362	4.76 $\times 10^{-5}$	6.03 $\times 10^{-4}$	0.09
Glucosamine-6-phosphate isomerase, conserved site	IPR018321	7.29 $\times 10^{-7}$	8.24 $\times 10^{-8}$	0.24
4-alpha-L-fucosyltransferase	IPR009993	9.36 $\times 10^{-5}$	8.82 $\times 10^{-5}$	0.19
Other transport and establishment of localization (2.62 $\times 10^{-6}$)				
GlpT transporter	IPR000849	2.04 $\times 10^{-11}$	1.45 $\times 10^{-18}$	0.31
Putative sugar-specific permease, SgaT/UlaA	IPR007333	1.25 $\times 10^{-11}$	5.18 $\times 10^{-16}$	0.33
ABC transporter, thiamine, ATP-binding protein	IPR005968	2.38 $\times 10^{-8}$	5.59 $\times 10^{-8}$	0.28
Type III secretion system needle protein	IPR011841	3.41 $\times 10^{-6}$	4.36 $\times 10^{-7}$	0.24
D-lactate dehydrogenase, membrane binding, C-terminal	IPR015409	1.71 $\times 10^{-6}$	4.34 $\times 10^{-6}$	0.22
Salmonella/Shigella invasion protein E	IPR003520	1.07 $\times 10^{-5}$	9.43 $\times 10^{-6}$	0.23

GO Biological process	**IPR entry**	**co-occurrence p-value**	**abundance p-value**	**correlation score**
Nucleoside:H+ symporter	IPR004740	9.40×10^{-5}	1.37×10^{-7}	0.19
Invasion protein B	IPR003065	8.09×10^{-5}	6.96×10^{-5}	0.21
Porin, LamB type	IPR003192	6.17×10^{-5}	3.81×10^{-11}	0.17
Nicotinamide mononucleotide transporter PnuC	IPR006419	4.53×10^{-5}	2.57×10^{-4}	0.15
Na-translocating NADH-quinone reductase subunit A	IPR008703	2.90×10^{-5}	1.03×10^{-3}	0.13
Other signal transduction (2.85×10^{-5})				
PhoQ Sensor	IPR015014	1.38×10^{-5}	1.29×10^{-5}	0.22
Other regulation of cellular process (1.61×10^{-2})				
S-ribosylhomocysteinase (LuxS)	IPR003815	3.04×10^{-8}	6.10×10^{-7}	0.26
PRD	IPR011608	3.01×10^{-7}	2.76×10^{-23}	0.16
Phage antitermination Q-like	IPR010534	4.13×10^{-6}	4.94×10^{-10}	0.20
Regulation modulator SeqA	IPR005621	7.15×10^{-5}	1.26×10^{-4}	0.18
Methionine repressor MetJ	IPR002084	4.67×10^{-5}	5.76×10^{-5}	0.19
Eukaryotic transcription factor, Skn-1-like, DNA-binding	IPR008917	1.43×10^{-5}	9.32×10^{-8}	0.20
Other biological regulation (6.98×10^{-3})				
Inhibitor of vertebrate lysozyme	IPR014453	4.50×10^{-7}	1.38×10^{-6}	0.23
CutC	IPR005627	1.09×10^{-6}	6.89×10^{-6}	0.22

Table 6.1: A list of mucosa-associated IPR domains annotated with overrepresented GO biological process.

Table 6.2: **A list of mucosa-associated IPR domains annotated with GO terms.** These GO biological process terms were not overrepresented in the GO term enrichment analysis of IPR domains with co-occurrence p-value < 1 $\times 10^{-2}$. The entries were sorted by the GO terms representing biological process and molecular function.

Description	**IPR entry**	**co-occurrence p-value**	**abundance p-value**	**correlation score**	**GO-Biological process**	**GO-Molecular function**
Citrate lyase, alpha subunit	IPR006472	9.26 $\times 10^{-5}$	3.24 $\times 10^{-4}$	0.16	acetyl-CoA metabolic process	citrate CoA-transferase activity
Aspartate–ammonia ligase	IPR004618	4.12 $\times 10^{-14}$	1.07 $\times 10^{-12}$	0.37	asparagine biosynthetic process	aspartate-ammonia ligase activity
Lysozyme subfamily 2	IPR013338	4.97 $\times 10^{-8}$	1.09 $\times 10^{-10}$	0.22	cell wall metabolic process	hydrolase activity
Ribonucleotide reductase	IPR000358	2.33 $\times 10^{-7}$	2.48 $\times 10^{-5}$	0.21	deoxyribonucleoside diphosphate metabolic process	ribonucleoside-diphosphate reductase activity
DNA mismatch repair protein MutH, conserved region	IPR018140	1.77 $\times 10^{-5}$	1.33 $\times 10^{-5}$	0.21	DNA modification	DNA binding, endonuclease activity
DNA polymerase III-theta, bacterial	IPR009052	1.38 $\times 10^{-5}$	1.58 $\times 10^{-7}$	0.21	DNA replication	DNA binding, DNA-directed DNA polymerase activity
DNA polymerase III, delta subunit, C-terminal	IPR015199	6.34 $\times 10^{-9}$	1.22 $\times 10^{-8}$	0.29	DNA replication	DNA binding, DNA-directed DNA polymerase activity
DNA polymerase III, psi subunit	IPR004615	3.72 $\times 10^{-5}$	4.38 $\times 10^{-5}$	0.20	DNA replication	DNA-directed DNA polymerase activity, 3’-5’ exonuclease activity
Ribonucleotide reductase N-terminal	IPR013554	4.11 $\times 10^{-7}$	2.84 $\times 10^{-6}$	0.22	DNA replication	ribonucleoside-diphosphate reductase activity, protein binding
DNA replication terminus site-binding protein	IPR008865	2.84 $\times 10^{-5}$	2.20 $\times 10^{-5}$	0.20	DNA replication termination	DNA binding
Fumarate reductase, D subunit	IPR003418	1.31 $\times 10^{-9}$	3.33 $\times 10^{-9}$	0.30	fumarate metabolic process	
Dihydrofolate reductase region	IPR001796	1.03 $\times 10^{-5}$	1.13 $\times 10^{-3}$	0.17	glycine biosynthetic process, nucleotide biosynthetic process	dihydrofolate reductase activity
Leucine operon leader peptide	IPR012570	2.05 $\times 10^{-5}$	1.77 $\times 10^{-5}$	0.22	leucine biosynthetic process	
Cof protein	IPR000150	1.17 $\times 10^{-6}$	2.55 $\times 10^{-22}$	0.28	metabolic process	hydrolase activity
Glycerate kinase	IPR004381	4.49 $\times 10^{-5}$	1.18 $\times 10^{-4}$	0.16	organic acid phosphorylation	glycerate kinase activity
Radical-activating enzyme, conserved site	IPR001989	3.56 $\times 10^{-6}$	5.83 $\times 10^{-14}$	0.22	oxygen and reactive oxygen species metabolic process	oxidoreductase activity, 4 iron, 4 sulfur cluster binding
Bordetella pertussis toxin A	IPR003898	8.09 $\times 10^{-5}$	6.96 $\times 10^{-5}$	0.21	pathogenesis	
Enterotoxin, bacterial	IPR008992	1.30 $\times 10^{-6}$	4.78 $\times 10^{-28}$	0.15	pathogenesis	
Invasion plasmid antigen IpaD	IPR009483	3.88 $\times 10^{-5}$	3.36 $\times 10^{-5}$	0.21	pathogenesis	

Description	IPR entry	co-occurrence p-value	abundance p-value	correlation score	GO-Biological process	GO-Molecular function
Type III secretion apparatus protein OrgA/MxiK	IPR013388	1.07×10^{-5}	9.43×10^{-6}	0.23	pathogenesis	
Glycoside hydrolase, family 25	IPR002053	6.18×10^{-6}	4.47×10^{-11}	0.19	peptidoglycan/cellwa catabolic process	lysozyme activity
Plasmid replication initiation, RepA	IPR003446	8.09×10^{-5}	1.77×10^{-5}	0.20	plasmid maintenance	
Peptidyl-prolyl cis-trans isomerase, FKBP-type, N-terminal	IPR000774	3.40×10^{-6}	2.79×10^{-5}	0.12	protein folding	
Peptidase C1B, bleomycin hydrolase	IPR004134	4.36×10^{-6}	9.81×10^{-13}	0.23	proteolysis	cysteine-type endopeptidase activity
Peptidase C69, dipeptidase A	IPR005322	1.62×10^{-5}	6.59×10^{-11}	0.18	proteolysis	dipeptidase activity
Peptidase S6, IgA endopeptidase	IPR000710	1.07×10^{-5}	9.32×10^{-8}	0.20	proteolysis	serine-type endopeptidase activity
Acid shock	IPR009435	1.41×10^{-6}	4.17×10^{-7}	0.24	response to acidity	
Cell division inhibitor SulA	IPR004596	4.12×10^{-5}	3.74×10^{-5}	0.21	SOS response	
Thr operon leader peptide	IPR011720	5.18×10^{-6}	4.51×10^{-6}	0.23	threonine biosynthetic process	
tRNA (guanine-N-7) methyltransferase	IPR003358	6.47×10^{-9}	5.02×10^{-4}	0.22	tRNA modification	tRNA (guanine-N7-)-methyltransferase activity

Table 6.2: A list of mucosa-associated IPR domains annotated with GO terms.

Table 6.3: **A list of selected mucosa-associated IPR domains without GO term annotation.** (A complete list is shown in Appendix F) Given a null hypothesis of no association between an InterPro (IPR) domain and mucosa-thriving microorganisms, a tests based on the hypergeometric distribution yielded significant p-values. Therefore, these IPR domains significantly co-occurred with the mucosa-thriving microorganisms.

Description	**IPR entry**	**co-occurrence p-value**	**abundance p-value**	**correlation score**
Uracil-DNA glycosylase, active site	IPR018085	3.16×10^{-12}	1.23×10^{-6}	0.30
Prokaryotic chromosome segregation and condensation protein MukF	IPR005582	2.90×10^{-10}	6.24×10^{-10}	0.32
dsDNA mimic, putative	IPR007376	8.75×10^{-10}	6.15×10^{-10}	0.31
Fumarate reductase, subunit C	IPR003510	1.31×10^{-9}	3.33×10^{-9}	0.30
Acid phosphatase (Class B)	IPR005519	2.12×10^{-9}	9.12×10^{-10}	0.29
NLPA lipoprotein	IPR004872	3.30×10^{-9}	5.00×10^{-9}	0.18
Tryptophan/tyrosine permease	IPR018227	7.29×10^{-9}	1.16×10^{-16}	0.23
Cyd operon protein YbgE	IPR011846	3.34×10^{-8}	4.78×10^{-8}	0.28
Mannitol repressor	IPR007761	3.34×10^{-8}	2.01×10^{-10}	0.27
Phosphomannose isomerase, type I, conserved site	IPR018050	3.34×10^{-7}	2.45×10^{-8}	0.26
Antimicrobial peptide resistance and lipid A acylation PagP	IPR009746	4.20×10^{-7}	1.10×10^{-6}	0.22
Porin, general diffusion Gram-negative, conserved site	IPR013793	4.50×10^{-7}	1.70×10^{-20}	0.24
C4-dicarboxylate anaerobic carrier-like	IPR018385	6.22×10^{-7}	6.93×10^{-13}	0.24
Tryptophan/tryrosine permease, conserved site	IPR013061	6.58×10^{-7}	2.50×10^{-12}	0.25
Adhesion, bacterial	IPR008966	2.08×10^{-6}	3.29×10^{-97}	0.23
Ionotropic glutamate receptor	IPR001320	8.61×10^{-6}	1.50×10^{-11}	0.21
Phosphotransferase system EIIB/cysteine phosphorylation site	IPR018113	9.07×10^{-6}	3.65×10^{-38}	0.25
DNA damage-inducible protein DinI-like	IPR010391	9.96×10^{-6}	1.58×10^{-10}	0.20
Amino acid transporter, transmembrane	IPR013057	1.08×10^{-5}	7.38×10^{-18}	0.08
Haemolysin expression modulating, HHA	IPR007985	1.38×10^{-5}	2.61×10^{-15}	0.23
Glycosyltransferase sugar-binding region containing DXD motif	IPR007577	1.50×10^{-5}	1.96×10^{-11}	0.20
Tetratricopeptide TPR-3	IPR011716	1.88×10^{-5}	6.12×10^{-11}	0.26
Chlamydia polymorphic membrane, middle domain	IPR011427	2.05×10^{-5}	1.27×10^{-69}	0.21
Phosphotransferase system, EIIC component, type 1	IPR013013	2.27×10^{-5}	1.43×10^{-36}	0.25
Pili assembly chaperone, conserved site	IPR018046	3.18×10^{-5}	1.02×10^{-32}	0.24
Opacity-associated protein A, N-terminal	IPR013731	3.08×10^{-5}	8.93×10^{-8}	0.21
FimH, mannose-binding	IPR015243	3.88×10^{-5}	7.23×10^{-7}	0.21
Mycoplasma MFS transporter	IPR011699	3.88×10^{-5}	7.23×10^{-7}	0.21
Glycosyl transferase, family 8	IPR002495	4.26×10^{-5}	9.27×10^{-20}	0.26
YidE/YbjL duplication	IPR006512	5.51×10^{-5}	1.46×10^{-10}	0.22
Prophage minor tail Z	IPR010633	8.09×10^{-5}	1.23×10^{-9}	0.17
Prophage tail fibre N-terminal	IPR013609	8.09×10^{-5}	8.58×10^{-14}	0.20
Phosphotransferase system, EIIC component, type 2	IPR013014	8.84×10^{-5}	5.71×10^{-26}	0.27
CblD like pilus biogenesis initiator	IPR010888	8.09×10^{-5}	6.96×10^{-5}	0.21
Secretion monitor	IPR009502	1.38×10^{-5}	1.29×10^{-5}	0.22

Table 6.3: A list of selected mucosa-associated IPR domains without GO term annotation.

Analysis of mucosal protein domains of secretomes and surface proteomes

Microbial cell surface proteins exist at the interface between the microorganism and the host mucosal environment. The surface proteome is an important factor in the survival strategy of microorganisms in the host body. To gain a better understanding of the functional perspective of the surface proteome and secretome of microorganisms thriving mucosal environments, the mucosa-associated IPR domains located on extracytoplasmic proteins were investigated further. To identify key features that are shared across a board range of mucosal microorganisms, the distribution of the putative extracytoplasmic mucosa-associated protein domains were also investigated. Out of 231 identified mucosa-associated IPR domains, more than 107 entries (47%) were found on extracytoplasmic proteins. Eighty-eight entries were found on 95-100% putative extracytoplasmic proteins of all sequences carrying that domain, whereas 19 entries found on 50-94% extracytoplasmic proteins (see Table 6.4 and Appendix H). The putative extracellular proteins were identified using the project identification workflow developed in this project (discussed in Chapter 4).

The results showed that most of the domains were presented in Bacterial members, and several are exclusive to a particular bacterial phylum. For example, the PhoQ Sensor (IPR015014) is unique to members of Proteobacteria-gamma.

For the set of strongly mucosa-associated domains, only two entries, IPR010619 and IPR008966, located on extracytoplasmic proteins were distributed across Archaea, Bacteria and Eukaryotes. The widely distributed mucosa-associated domains, suggesting the importance of these domains for the survival of microorganisms in mucosal environments. The former entry has not yet been characterised. It might be worth carrying out a further detailed investigation of the involvement of this conserved region in the context of mucosa-microbe interactions. The latter widely distributed surface mucosa-associated domain (IPR008966) was characterised as a bacterial adhesin.

Table 6.4: **A summary list of mucosa-associated protein domains located on extracytoplasmic proteins.** (see Appendix H for a complete list)

IPR	description	Dist.	Clas dist.	Total	(%) extprot	PROTI	FUN	ARC	ACI	ACT	AQU	BAC	CHLA	CHLO	CHLOF	CYA	DIC	ELU	FIR	FUS	NIT	PLA	PRO	SPI	TEN	THEMI	THEMO	VER
IPR010619	Protein of unknown function DUF1212	ABE	18	433	99.31	M	AB	MbMm		X		X				X				X		X	ABDEG	X				X
IPR006512	YidE/YbjL duplication	AB	16	293	100.00			H	X	X		X			X	X			X	X		X	ABDEG				X	X
IPR004445	Sodium/glutamate symporter	AB	15	266	100.00			Mm	X	X		X				X			X	X		X	ABDEG	X			X	
IPR002053	Glycoside hydrolase, family 25	BE	15	404	67.33	EMP	AB			X		X			X	X			X				ADEG	X				
IPR004872	NLPA lipoprotein	AB	14	929	96.66			A		X		X	X						X	X			ABDEG	X		X	X	
IPR013014	Phosphotransferase system, EIIC component, type 2	AB	13	978	100.00			H		X		X			X				X	X			ABG	X	X	X	X	
IPR005185	Protein of unknown function DUF307	BE	13	209	100.00	MP	AB			X		X		X		X			X				ABDG					
IPR013011	Phosphotransferase system, EIIB component, type 2	AB	13	1612	62.84			H		X					X		X		X	X			ABG	X	X	X	X	
IPR018385	C4-dicarboxylate anaerobic carrier-like	B	12	398	100.00					X		X						X	X	X			ABEG	X	X		X	
IPR006419	Nicotinamide mononucleotide transporter PnuC	B	12	348	100.00					X		X			X	X			X				ABDEG	X				X
IPR013057	Amino acid transporter, transmembrane	BE	12	499	100.00	ADEMPU	ABM			X									X				G					
IPR000774	Peptidyl-prolyl cis-trans isomerase, FKBP-type, N-terminal	BE	12	520	70.96	A			X			X	X									X	ABDEG	X				X
IPR001127	Phosphotransferase system, sugar-specific permease EIIA 1 domain	AB	12	840	58.81			Mm	X									X	X	X			ABEG	X	X	X		
IPR001320	Ionotropic glutamate receptor	B	11	425	100.00					X					X	X			X				ABDEG			X	X	
IPR013013	Phosphotransferase system, EIIC component, type 1	B	11	1401	100.00					X								X	X	X			ABEG	X	X	X		
IPR018227	Tryptophan/tyrosine permease	AB	11	484	100.00	E		Tc		X			X			X			X				ABDEG					
IPR018113	Phosphotransferase system EIIB/cysteine phosphorylation site	B	11	1409	97.44					X								X	X	X			ABEG	X	X	X		
IPR001996	Phosphotransferase system, EIIB	B	11	1452	95.94					X								X	X	X			ABEG	X	X	X		
IPR005519	Acid phosphatase (Class B)	BE	11	216	87.96		B			X		X		X		X			X				AEG		X			X
IPR004685	Branched-chain amino acid transport system II carrier protein	B	10	483	100.00					X			X						X	X			ABDEG	X				
IPR008966	Adhesion, bacterial	ABE	10	2333	95.71		A	MbMm		X					X				X			X	BDG					

IPR	description	Dist.	Clas dist.	Total	(%) extprot	PROTI	FUN	ARC	ACI	ACT	AQU	BAC	CHLA	CHLO	CHLOF	CYA	DIC	ELU	FIR	FUS	NIT	PLA	PRO	SPI	TEN	THEMI	THEMO	VER
IPR003501	Phosphotransferase system, lactose/cellobiose-specific IIB subunit	B	9	1186	64.42					X					X		X		X				ABG	X	X			
IPR008142	Alanine dehydrogenase/pyridine nucleotide transhydrogenase, conserved site-1	BE	9	222	59.91		A		X	X		X				X			X				ABG					
IPR000849	GlpT transporter	B	8	337	100.00					X			X						X				ABEG	X				
IPR004740	Nucleoside:H+ symporter	B	7	164	100.00				X	X		X						X				X	AG					
IPR004704	Phosphotransferase system, mannose/fructose/sorbose family IID component	AB	7	473	99.58			McTp										X	X	X			DG					
IPR013338	Lysozyme subfamily 2	B	7	355	65.92					X		X							X				ABG				X	
IPR004703	Phosphotransferase system, galactitol-specific IIC component	B	6	193	100.00					X		X							X				ABG					
IPR004700	Phosphotransferase system, sorbose-specific IIC subunit	AB	6	462	100.00			Tp										X	X	X			DG					
IPR004699	Phosphotransferase system, enzyme II sorbitol-specific factor	B	6	91	98.90					X					X				X				ABG					
IPR011638	Sorbitol phosphotransferase enzyme II, C-terminal	B	6	90	98.89					X					X				X				ABG					
IPR009693	Glucitol operon activator	B	6	88	98.86					X					X				X				ABG					
IPR003192	Porin, LamB type	B	5	191	96.34				X														ABDG					
IPR009993	4-alpha-L-fucosyltransferase	B	4	79	100.00																		BEG				X	
IPR013012	Phosphotransferase system, EIIB component, type 3	B	4	455	83.74					X									X				G	X				
IPR013061	Tryptophan/tryrosine permease, conserved site	B	3	205	100.00																		BDG					
IPR008992	Enterotoxin, bacterial	B	3	314	99.36														X				BG					
IPR010486	HNS-dependent expression A	B	3	44	97.73																		ABG					
IPR014453	Inhibitor of vertebrate lysozyme	B	3	81	97.53																		ABG					
IPR000710	Peptidase S6, IgA endopeptidase	B	3	53	79.25																		BEG					
IPR005968	ABC transporter, thiamine, ATP-binding protein	B	2	116	100.00																		AG					
IPR013793	Porin, general diffusion Gram-negative, conserved site	B	2	288	100.00																		BG					
IPR018046	Pili assembly chaperone, conserved site	B	2	654	100.00																		BG					

IPR	description	Dist.	Clas dist.	Total	(%) extprot	PROTI	FUN	ARC	ACI	ACT	AQU	BAC	CHLA	CHLO	CHLOF	CYA	DIC	ELU	FIR	FUS	NIT	PLA	PRO	SPI	TEN	THEMI	THEMO	VER
IPR009746	Antimicrobial peptide resistance and lipid A acylation PagP	B	2	78	98.72																		BG					
IPR009435	Acid shock	B	2	66	98.48																		BG					
IPR010888	CblD like pilus biogenesis initiator	B	2	29	86.21																		BG					
IPR014318	Phage shock protein G	B	1	66	100.00																		G					
IPR010771	Intracellular growth attenuator IgaA	B	1	68	100.00																		G					
IPR006817	LPP motif	B	1	81	100.00																		G					
IPR015014	PhoQ Sensor	B	1	68	100.00																		G					
IPR011427	Chlamydia polymorphic membrane, middle domain	B	1	174	100.00								X															
IPR003517	Cysteine-rich outer membrane protein 3, Chlamydia	B	1	12	100.00								X															
IPR000604	Major outer membrane protein, Chlamydia	B	1	12	100.00								X															
IPR011699	Mycoplasma MFS transporter	B	1	17	100.00																				X			
IPR010794	Maltose operon periplasmic	B	1	83	98.80																		G					
IPR015243	FimH, mannose-binding	B	1	60	95.00																		G					
IPR009502	Secretion monitor	B	1	63	76.19																		G					

Table 6.4: A summary list of mucosa-associated protein domains located on extracytoplasmic proteins. (see Appendix H for a complete list) The mucosa-associated domains identified using co-occurrence p-value cut-off of < 1E-04. A domain was indicated as located on extracytoplasmic protein if more than 50% of all proteins (that were included in this study) carrying that domain were predicted as extracytoplasmic protein by our sequence analysis pipeline. 'Dist.' denotes the distribution of the given domain across superkingdom where A=Archaea, B=Bacteria, E=Eukaryote. 'Class dist.' represents number of taxonomic classification that the domain was annotated. The taxonomic classification system used here are denoted as the sideway headers. 'Total' indicates number of protein sequences predicted to carry the domain. '(%) extprot' represents the proportion of domain-contains sequences that were predicted as extracytoplasmic proteins. Taxonomic classification: PROT=Protist where A=Apicomplexa, D=Diplomonadida, E=Entamoebidae, U=Euglenozoa, M=Mycetozoa, and P=Parabasalidea. FUN=Fungi where A=Ascomycota, B=Basidomycota, and M=Microsporidia. ARC=Archaea where A=Archaeoglobi, H=Halobacteria, Mb=Methanobacteria, Mm=Methanomicrobia, Tc=Thermococci, and Tp=Thermoplasma. ACI=acidobacteria, ACT=Actinobacteria, AQU=Aquificae, Bac=Bacteroidetes, CHLA=Chlamydiae, CHLO=Chlorobi, CHLOF=Chloroflexi, CYA=Cyanobacteria, DIC=Dictyoclomi, ELU=Elusimicrobia, FIR=Firmicutes, FUS=Fusobacteria, NIT=Nitrospirae, PLA=Plancotomycetes, PRO=Proteobacteria where A=PRO-alpha, B=PRO-beta, D=PRO-delta, E=PRO-epsilon, and G=PRO-gamma. SPI=Spirochaetes, TEN=Tenericutes, THEMI=Thermi, THERMO=Thermotogae, VER=verrucomicrobia.

6.3.3 Clustering analysis of extracytoplasmic proteins of mucosa-thriving microbes

As a result of clustering putative extracytoplasmic proteins from a set of 75 mucosa-thriving microorganisms (see Table 6.5), 8,895 clusters of similar protein sequences were identified with a BLAST e-value of $< 1 \times 10^{-5}$ and percent identity of > 50. Out of all 82,863 extracytoplasmic proteins, 73,686 were grouped into 8,895 clusters of either paralogous or orthologous protein pairs, leaving the remaining 9,177 sequences unclustered. The biggest protein family has a membership of 746 paralogous proteins from *T. vaginalis*. All of these *T. vaginalis*' paralogous proteins are annotated as hypothetical proteins with no known function. Highly similar sequences derived from the same taxa were considered as paralogs, whereas orthologs were classed as homologous sequences from different taxa. About half of the protein families contained at least one member regarded as a protein overrepresented in mucosal organisms. Mucosa-specific proteins were identified by similarity searches (using a BLASTP search with cut-off e-value of 1×10^{-5}) resulting in hits to a significant number of known mucosa-associated organisms (at the p-value cut-off of 1×10^{-2}) (see method Section 6.2.2).

Approximately 85% (7538/8895) of the protein families comprised of members from a single taxonomic classification (see Table 6.6). Seventy-five taxa included in the clustering analysis were from eleven different taxonomic classes according to the GOLD database taxonomic classification (see Figure 6.11). These eleven taxonomic classes included were Chlamydiae, Fusobacteria, Firmicutes, Actinobacteria, Bacteroidetes, Epsilon-Proteobacteria, Gamma-Proteobacteria, Apicomplexa, Entamoebidae, Parabasalidea, and Diplomonadida. Bacteroidetes proteins appeared to have the highest number of single-class clusters (singleton), followed by those from Gamma-Proteobacteria, Entamoebidae, Firmicutes and Parabasalidea, respectively (see Figure 6.12). These families represent the exclusive set of extracytoplasmic proteins within the organisms' groups that are specific to particular groups of mucosal microbial communities. These results suggest that there is a high degree of variation across the Bacteroidetes group-specific families (see Appendix E). This high degree of diversity might result from the more diverse subgroups, at the 'Genus' level of the Bacteroidetes data set in contrast to other taxonomic classes. Moreover, this variation might reflects the different adaptation of the Bacteroidetes to specific conditions or niches [Sonnenburg *et al.*, 2010]. The families from the Gamma-Proteobacteria revealed a large number of different homologous groups even though the proteins were taken from various strains of one specie (e.g. *Escherichia coli*). The diversity of the *E. coli* extracytoplasmic protein family is notably greater than those of the Firmicutes class (see Figure 6.12), which included various species from the Genus Lactobacillus. This result suggests that the high level of variation seen in the extracytoplasmic proteomes of the *E. coli* strains might reflect the

Table 6.5: Summary of the number of proteins from the clustering analysis of extracytoplasmic proteins from 75 known mucosa-thriving microorganisms.

Number of proteins analysed (from 75 mucosa-thriving microbes)	285,047
Number of putative extracytoplasmic proteins	82,863
of which:	
significantly associated with mucosa organisms (p-value < 1×10^{-2})	37,070
Number of extracytoplasmic protein cluster after clustering	8,895 (73,686 proteins)
of which:	
contain at least one protein significantly associated with mucosa microbes	4,377
no protein signatures detected by InterProScan	2,735
Number of clusters without domain annotation and contains at least one protein significantly associated with mucosa microbes	916
Number of extracytopasmic proteins not clustered	9,177

Table 6.6: The number of extracytoplasmic protein cluster counted based on the number of taxonomic class presented. The taxonomic classes used here were obtained from the GOLD database taxonomic classification. The 75 taxa included in the clustering analysis are originated from eleven taxonomic classes. None of the clusters are shared among all 11 classes. Not surprisingly, most of the proteins clusters are shared within one or two classes.

Number of taxonomic class presented in the cluster	**Number of clusters**
1	7538
2	769
3	307
4	158
5	62
6	37
7	17
8	3
9	2
10	2
Total	8895

ability of these microbes to thrive in diverse host environments. The *E. coli* strains included in the data set are known commensals or pathogens in many mucosa-lined niche environments such as the human intestine and urogenital tract, as well as avian lungs [Kaper, 2005][Rasko *et al.*, 2008].

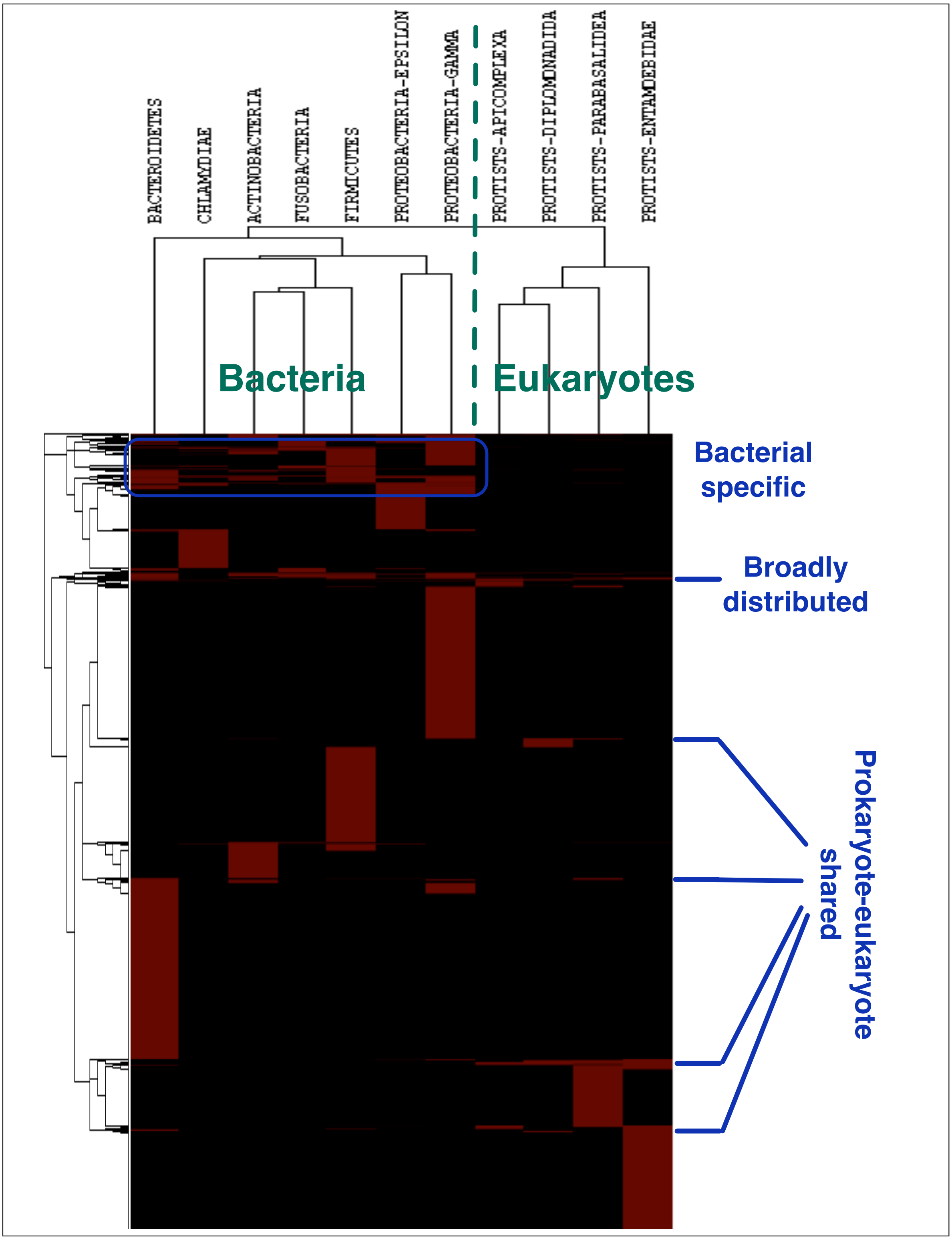


Figure 6.11: Dendrogram of 8,895 extracytoplasmic protein families across 11 taxonomic classes. The protein clusters were grouped based on their distribution pattern across the taxonomic classification. The outstanding patterns of interest are annotated in the blue text. Most of the families were derived from within a taxonomic class. Only a small proportion of the families are distributed across the bacterial taxonomic classes, and even no family contain members from all 11 taxonomic groups. The maximum number of classes found distributed in two clusters are 10.

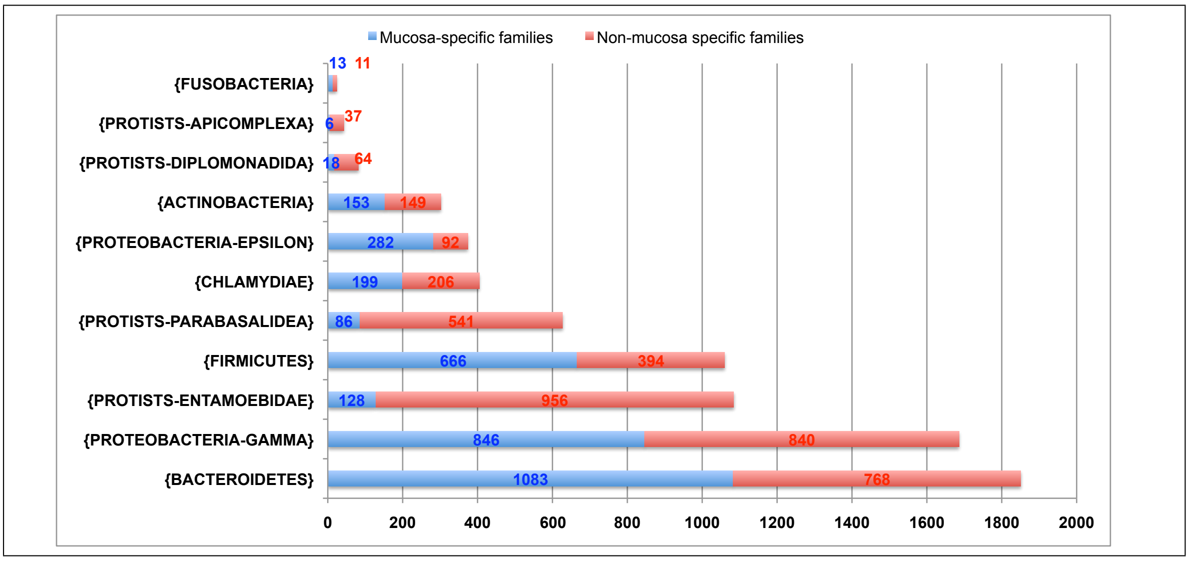


Figure 6.12: The number of extracytoplasmic protein families exclusive to a particular taxonomic classes. Each family was evaluated for being a protein from a mucosa-associated family. A family was considered as mucosa-associated if at least one member of that family was identified as mucosa-specific protein. A protein sequence was determined whether it is mucosa-specific based on the BLASTP search against RefSeq (see Section 6.2.2 for more detail).

Not surprisingly, most of clusters with pair taxonomic classes are either among bacteria or among microbial eukaryotes (see Figure 6.13). Several clusters were shared between a pair of members of the prominent gut bacterial phyla communities (Bacteroidetes, Firmicutes, Actinobacteria, Proteobacteria). Some of the paired-class clusters contained members from both prokaryotes and eukaryotic microbes, suggesting lateral gene transfers between microorganisms sharing the same niches.

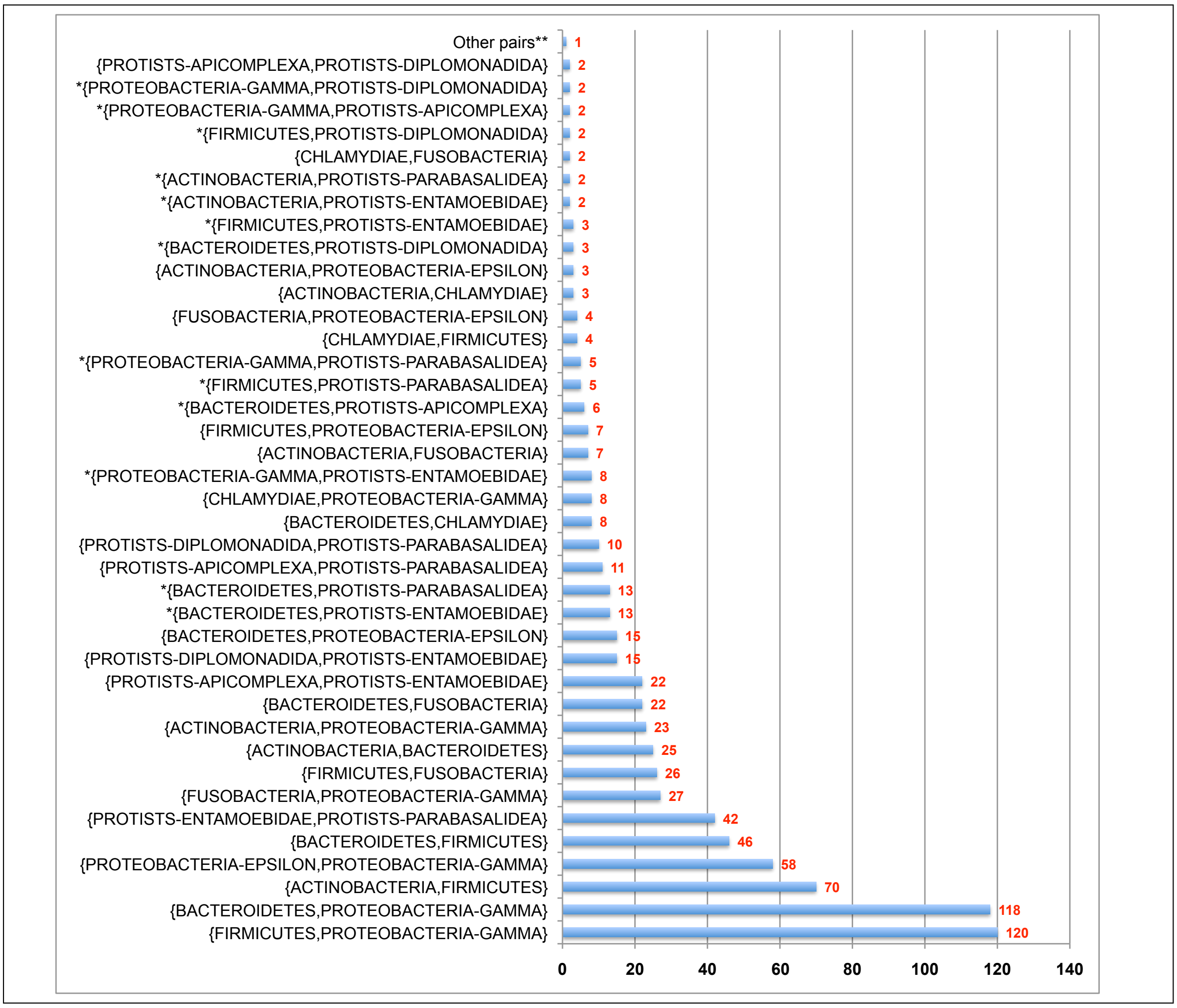


Figure 6.13: The number of extracytoplasmic protein families exclusive to two particular taxonomic classes. '*' denotes pairs of eukaryotic and prokaryotic classes. '**' represents [FIRMICUTES,PROTISTS-APICOMPLEXA], [PROTEOBACTERIA-EPSILON,PROTISTS-ENTAMOEBIDAE], [CHLAMYDIAE,PROTEOBACTERIA-EPSILON], [CHLAMYDIAE,PROTISTS-APICOMPLEXA], [CHLAMYDIAE,PROTISTS-DIPLOMONADIDA]

Twenty-four clusters appeared to comprise of members from all four protist classes (Aplicomplexa, Diplomonadida, Entamoebidae, and Parabasalidea) (see Figure 6.14). A lower proportion of families were shared across several bacterial and the eukaryotic classes compared to families of shared either within bacteria or eukaryotes (see Figure 6.11).

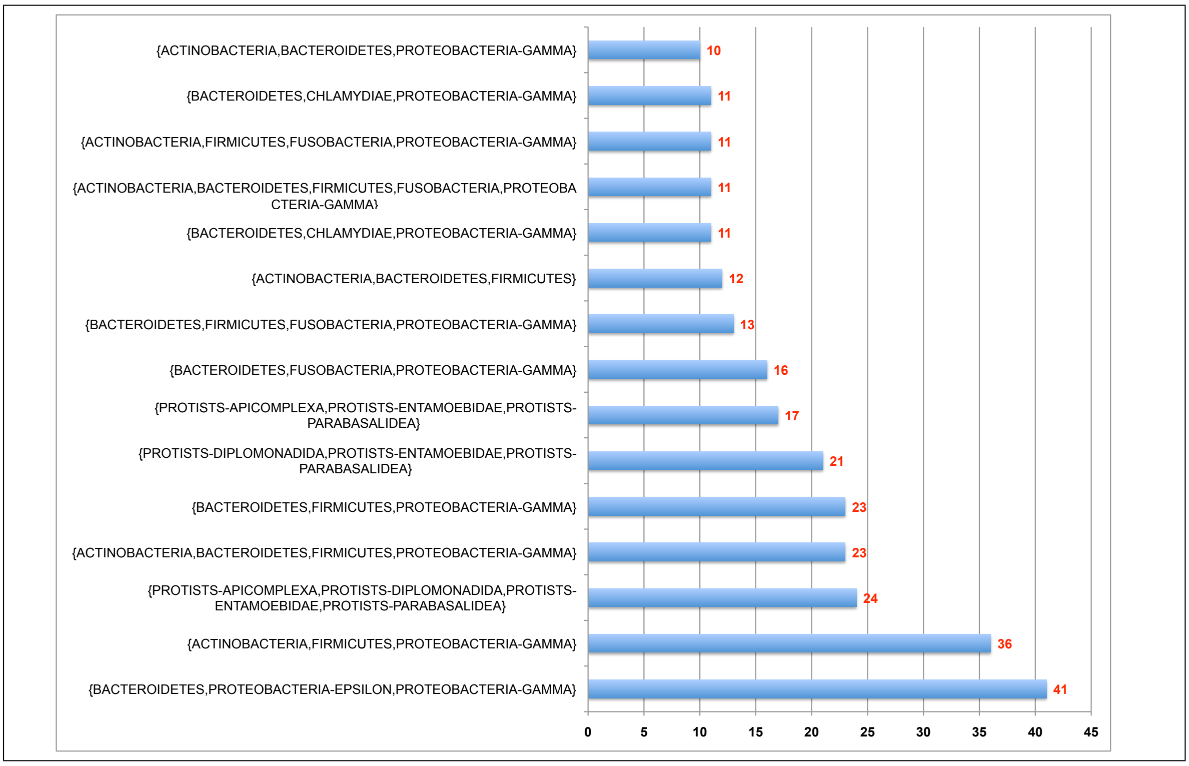


Figure 6.14: The number of extracytoplasmic protein families with shared members from more than two taxonomic classes. This chart only represents the distribution patterns of taxonomic classes for which more than nine clusters were found.

Cluster of widely distributed extracytoplasmic proteins

None of the clusters (see method Section 6.2.2) were shared across 75 known mucosa-thriving microbes included in the analysis. Two protein clusters (cluster_4 and cluster_67) that have the widest distribution of organisms contained proteins from 10 of the 11 taxonomic classes (see Figure 6.15). Cluster_4 contains 165 protein sequences from all classes except Diplomonadida (*Giardia lamblia* ATCC 50803). Among these 165 sequences, 78 (47.3%) were identified as mucosa-associated proteins (p-value < 1 $\times 10^{-2}$) by a significant proportion of BLASTP hits (against RefSeq; e-value < 1 $\times 10^{-5}$) to protein members of known mucosa-associated organisms. Based on the sequence similarity search against sequences in the COG database (see Method, Section 6.2.2), all members of this family were highly similar to proteins in COG1132 (ATPase and permease components of ABC-type multidrug transport system) and KOG0256 (1-aminocyclopropane-1-carboxylate synthase) with the best BLAST hit e-value < 1 $\times 10^{-10}$. Both COG and KOG families are annotated to be involved in defence and signal transduction mechanisms, respectively.

Cluster_67 contains 55 members, of which 18 (55%) were identified as mucosa-associated, from all classes except from Fusobacteria. Using the same approach to assign function to the cluster, the cluster was annotated with COG1028, COG0300, COG4221 and KOG4367 (see Figure 6.15).

These clusters of orthologs are poorly characterised. The other five protein clusters were found to share proteins from 8-9 taxonomic classes. These broadly distributed extracytoplasmic protein families were part of transporter and metabolisms of inorganic ion, lipid and coenzymes as well as post-translational modification, protein turnover and chaperones.

Protein clusters shared across prokaryotes and eukaryotes

Lateral gene transfer (LGT) allows organisms to incorporate genetic materials from taxa that are not their direct ancestors. Genes acquired through LGT may facilitate the adaptation of organisms to survive in a certain environment [Bellgard *et al.*, 2009]. For example, several antibiotics resistance genes are proposed to be transferred horizontally among gut microbiota [Salyers *et al.*, 2004]. The sequence identity of more than 99% is shown among some of these resistance genes from different species of both Gram-positive and Gram-negative bacteria [Gupta *et al.*, 2003]. Hence, LGT may contribute to the adaptation of microorganisms to survive in a specific niche [Bellgard *et al.*, 2009] [Xu *et al.*, 2007].

A considerable number of LGTs have been inferred to have taken place among prokaryotes as well as unicellular eukaryotes and across domains of life [Dutta and Pan, 2002][Keeling and Palmer, 2008]. Therefore, genes encoding extracytoplasmic proteins that are distributed across known mucosa-thriving prokaryotes and microbial eukaryotes might be a result of LGT. Such proteins could be involved in critical mechanisms for the survival of the microbes in the host mucosal environment.

Several extracytoplasmic protein clusters contain at least one protein from a known mucosa-thriving prokaryote as well as a mucosa-thriving microbial eukaryote (see Figure 6.16). Functional annotation based on COGs/KOGs revealed that most of the protein members these prokaryote-eukaryote clusters involve in transports and metabolisms of carbohydrate, inorganic ion and amino acids. Several others have no known specific functions. A few others were annotated as proteins involving in defense mechanisms, cell wall and membrane biogenesis, transcription and signal transduction.

Sets of protein clusters with no known protein domains

To reveal potentially important functional regions required by host-microbe interactions that have not yet identified, the extracytoplasmic protein families of the 75 known mucosa-thriving microorganisms were examined for families with no known protein signatures. Families whose members were not annotated with any known protein signatures can be considered to be clusters of proteins with potential new conserved regions.

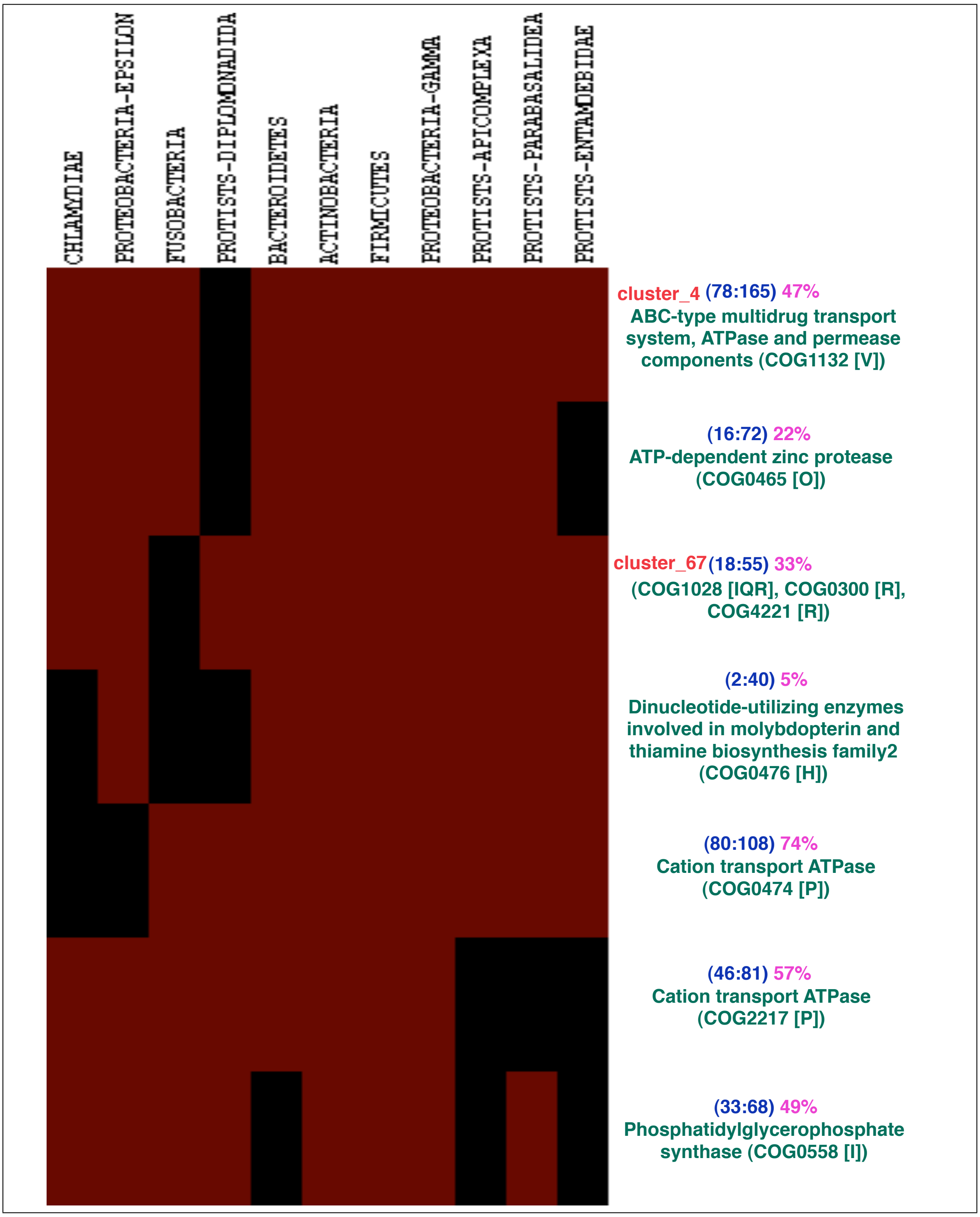


Figure 6.15: Heatmap of the seven protein clusters with the widest taxonomic distribution. For each entry, a description is given and the COG number is given. The letters in the square brackets refer to the functional categories as defined in the COG database: [P]=Inorganic ion transport and metabolism, [O]=Posttranslational modification, protein turnover, chaperones, [R]=General function prediction only, [H]=Coenzyme transport and metabolism ,[I]=Lipid transport and metabolism ,[V]=Defense mechanisms. COG1028 = Dehydrogenases with different specificities related to short-chain alcohol dehydrogenases, COG0300 = Short-chain dehydrogenases of various substrate specificities, COG4221 = Short-chain alcohol dehydrogenase of unknown specificity.

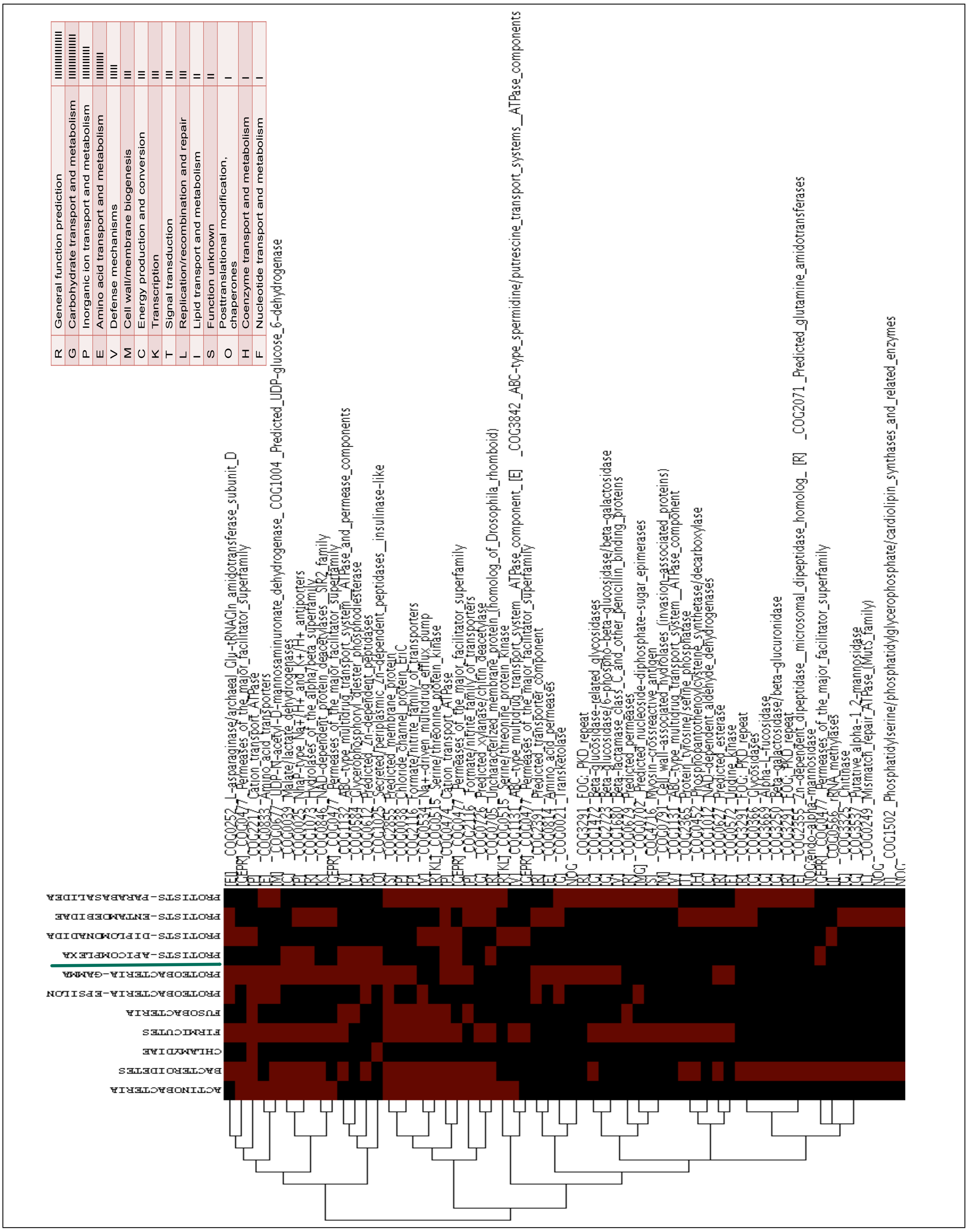


Figure 6.16: Dendrogram representing clusters of orthologous proteins shared between prokaryotes and microbial eukaryotes that are known to thrive in mucosal environments. The eleven taxonomic classes are shown. Dendrogram of protein clusters were clustered based on their taxonomic distribution. More than 50% of protein members of each of the clusters listed here were identified to be overrepresented among mucosal microorganisms (see 6.2.2) . For each COG entry, a description is given and the COG number is given. The letters in the square brackets refer to the functional categories as defined in the COG database (see Appendix M). NOC = not on COG. NOC:endo-alpha-mannosidase was assigned to one of the cluster due to the annotation of a protein from the Bacteroidetes.

Table 6.7: Summary of the number of mucosa-associated extracytoplasmic protein families that have no known protein domain in relation to their taxonomic class distribution. '*' denotes a shared cluster between members of microbial eukaryotic and prokaryotic classes.

distribution	Number cluster
BACTEROIDETES	257
FIRMICUTES	195
PROTEOBACTERIA-GAMMA	166
PROTEOBACTERIA-EPSILON	110
CHLAMYDIAE	72
ACTINOBACTERIA	63
PROTISTS-ENTAMOEBIDAE	14
PROTISTS-PARABASALIDEA	12
FIRMICUTES, FUSOBACTERIA	5
ACTINOBACTERIA, FIRMICUTES	4
BACTEROIDETES, FIRMICUTES	4
BACTEROIDETES, PROTEOBACTERIA-GAMMA	4
FUSOBACTERIA	2
ACTINOBACTERIA, BACTEROIDETES	1
BACTEROIDETES, FIRMICUTES, FUSOBACTERIA	1
*BACTEROIDETES, PROTISTS-PARABASALIDEA (endo-alpha-mannosidase)	1
FIRMICUTES, PROTEOBACTERIA-EPSILON	1
FIRMICUTES, PROTEOBACTERIA-GAMMA	1
PROTISTS-APICOMPLEXA, PROTISTS-ENTAMOEBIDAE	1
PROTISTS-DIPLOMONADIDA, PROTISTS-ENTAMOEBIDAE, PROTISTS-PARABASALIDEA	1
PROTISTS-ENTAMOEBIDAE, PROTISTS-PARABASALIDEA	1
Total	916

Among the 8,895 identified extracytoplasmic protein families, one third (2,735) of the families consisted solely of members that do not possess any known protein domains or signatures (using InterProScan search excluding regions of low complexity segment). Among the 2,735 clusters, 916 contained at least one protein sequence regarded as significantly mucosa-associated (see Table 6.5). Interestingly, most of these so called mucosa-specific uncharacterised protein families are sequences from one individual taxonomic class of the prominent human gut microbiome including Bacteroidetes, Firmicutes, Proteobacteria. These taxonomic classes appeared to have relatively high numbers of clusters of no known protein signatures (110-257 clusters) (see Table 6.7).

Several of the single-class protein clusters had no members possessing any known protein signatures, suggesting how little is known about the key components or mechanisms involved in the host-microbe interaction among each particular taxonomic class of naturally mucosa-thriving microbes.

Among the mucosa-specific uncharacterised extracytoplasmic protein families of multiple taxonomic classes, one particular family stood out because the members are common to the gut bacterial commensals *B. thetaiotaomicron VPI-5482* and the urogenital tract parasite *Trichomonas vaginalis G3* (see Table 6.8). This family contains 4 protein homologs, 3 paralogs from *T. vaginalis* and one from the *Bacteroides* specie. Interestingly, even though no characterised conserved InterPro domains were found, the Bacteroidetes gene product was annotated as an endo-alpha-mannosidase.

Table 6.8: A list of extracytoplasmic mucosa-associated protein families with no known domains. Members of these clusters were derived from at least 2 taxonomic classes. ‘*’ denotes a commonality between members of microbial eukaryotic and prokaryotic classes.

clusterid	distribution	class count	taxon count	gi count
3125	BACTEROIDETES,FIRMICUTES,FUSOBACTERIA	3	6	6
4452	PROTISTS-DIPLOMONADIDA,PROTISTS-ENTAMOEBIDAE,PROTISTS-PARABASALIDEA	3	4	4
409	BACTEROIDETES, PROTEOBACTERIA-GAMMA	2	26	28
509	BACTEROIDETES, PROTEOBACTERIA-GAMMA	2	25	25
1074	FIRMICUTES, PROTEOBACTERIA-GAMMA	2	22	22
2046	FIRMICUTES, FUSOBACTERIA	2	10	10
2193	FIRMICUTES, FUSOBACTERIA	2	9	9
2410	ACTINOBACTERIA, FIRMICUTES	2	8	8
3148	BACTEROIDETES, FIRMICUTES	2	6	6
3244	PROTISTS-ENTAMOEBIDAE, PROTISTS-PARABASALIDEA	2	3	5
3592	FIRMICUTES, FUSOBACTERIA	2	5	5
3856	*BACTEROIDETES, PROTISTS-PARABASALIDEA	2	2	4
3895	FIRMICUTES, FUSOBACTERIA	2	3	4
4040	FIRMICUTES, FUSOBACTERIA	2	4	4
4228	FIRMICUTES, PROTEOBACTERIA-EPSILON	2	4	4
4435	BACTEROIDETES, FIRMICUTES	2	4	4
4695	ACTINOBACTERIA, FIRMICUTES	2	2	3
4954	ACTINOBACTERIA, FIRMICUTES	2	3	3
5050	BACTEROIDETES, FIRMICUTES	2	3	3
5368	BACTEROIDETES, FIRMICUTES	2	3	3
5418	PROTISTS-APICOMPLEXA, PROTISTS-ENTAMOEBIDAE	2	3	3
6786	BACTEROIDETES, PROTEOBACTERIA-GAMMA	2	2	2
8029	ACTINOBACTERIA, BACTEROIDETES	2	2	2
8278	ACTINOBACTERIA, FIRMICUTES	2	2	2
8474	BACTEROIDETES, PROTEOBACTERIA-GAMMA	2	2	2

6.4 Discussion

In this chapter, the aim was to identify genotypic features (protein domains and clusters) that are overrepresented across mucosa-thriving microorganisms. Such features might contribute to specific adaptations of the microbes to the mucosal environments [Sonnenburg *et al.*, 2010]. In particular, these analyses focused on microbial extracytoplasmic proteins that are known to play important roles in the host-microbe and microbe-microbe interactions. Several protein domains and clusters were identified in this study that are overrepresented among mucosal microorganisms.

Two types of functional annotation were used in order to gain insight into the potential functional relevance of identified protein domains and clusters. GO terms associated to protein domains were used to identify the function categories enriched across mucosal-associated domains. In addition, COG and KOG functional annotations were used to assign possible functions to mucosa-associated protein clusters. The analysis of proteins domains provides the identification of potential mucosa-associated known conserved protein regions. The domain-based analysis was complemented with the protein clustering analysis. The clustering analysis allows the identification of potential new protein domains from a set of mucosa-associated clusters of homologous extracytoplasmic proteins.

Figure 6.17: **Heatmap showing the distribution of extracytoplasmic mucosa-associated protein domains across an organisms' origin of isolation.** Domains listed (vertical axis) are overrepresented among mucosa-associated taxa. Different habitat groups are listed on the horizontal axis. Both axes are grouped according to the distribution pattern. Interesting, several domains are encoded by different taxa from various mucosal environments. For example, the domains involving in phosphotransferase system (PTS) are presented in taxa known as human gastrointestinal, respiratory and urogenital tracts. 'git'=gastrointestial tract, 'ugt'=urogenital tract and 'rt'=respiratory tract.

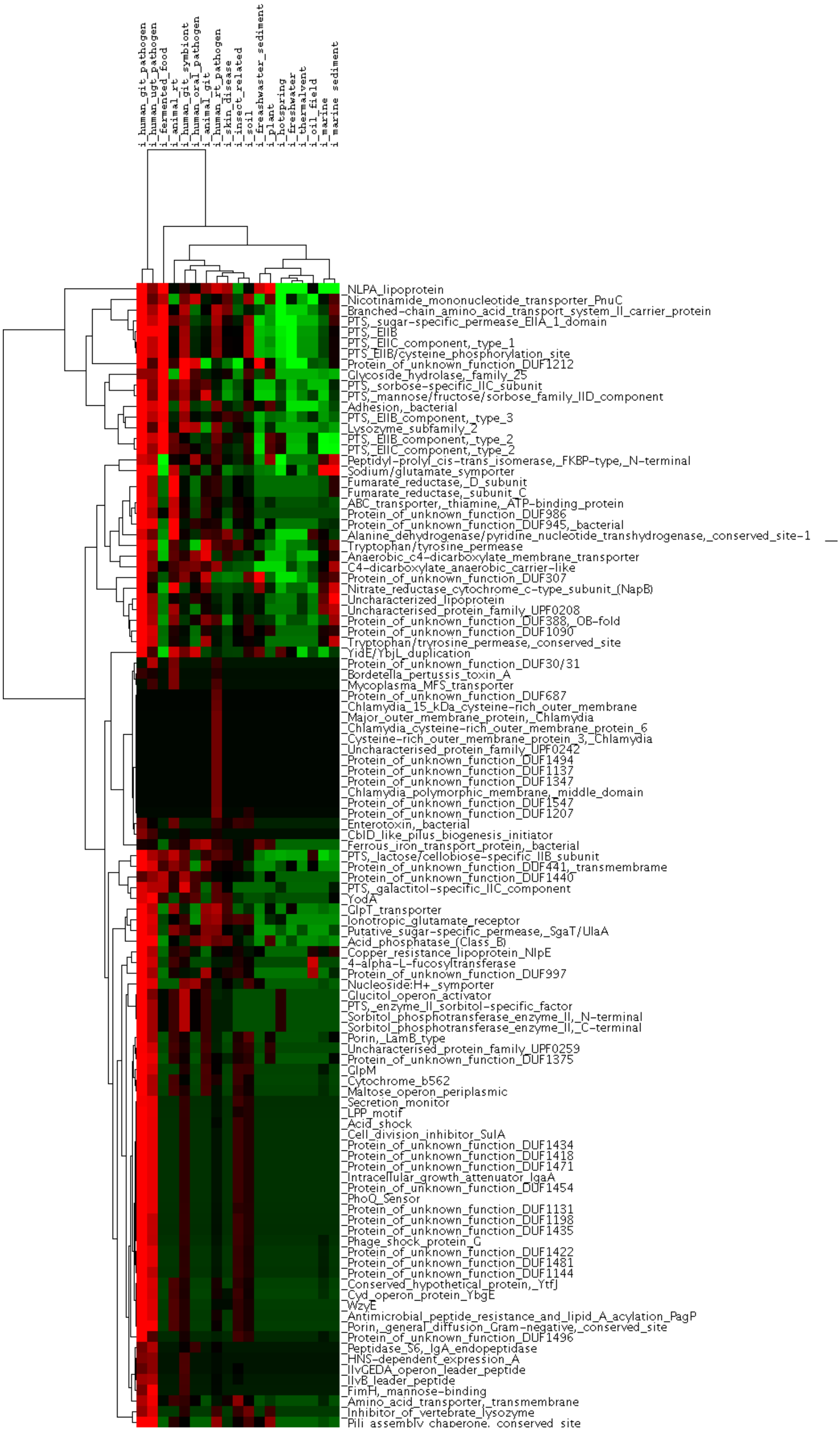

i_human_git_pathogen
i_human_ugt_pathogen
i_fermented_food
i_animal_rt
i_human_git_symbiont
i_human_oral_pathogen
i_animal_git
i_human_rt_pathogen
i_skin_disease
i_insect_related
i_soil
i_freashwaster_sediment
i_plant
i_hotspring
i_freshwater
i_thermalvent
i_oil_field
i_marine
i_marine sediment
NLPA_lipoprotein
Nicotinamide_mononucleotide_transporter_PnuC
Branched-chain_amino_acid_transport_system_II_carrier_protein
PTS,_sugar-specific_permease_EIIA_1_domain
PTS,_EIIB
PTS,_EIIC_component,_type_1
PTS_EIIB/cysteine_phosphorylation_site
Protein_of_unknown_function_DUF1212
Glycoside_hydrolase,_family_25
PTS,_sorbose-specific_IIC_subunit
PTS,_mannose/fructose/sorbose_family_IID_component
Adhesion,_bacterial
PTS,_EIIB_component,_type_3
Lysozyme_subfamily_2
PTS,_EIIB_component,_type_2
PTS,_EIIC_component,_type_2
Peptidyl-prolyl_cis-trans_isomerase,_FKBP-type,_N-terminal
Sodium/glutamate_symporter
Fumarate_reductase,_D_subunit
Fumarate_reductase,_subunit_C
ABC_transporter,_thiamine,_ATP-binding_protein
Protein_of_unknown_function_DUF986
Protein_of_unknown_function_DUF945,_bacterial
Alanine_dehydrogenase/pyridine_nucleotide_transhydrogenase,_conserved_site-1
Tryptophan/tyrosine_permease
Anaerobic_c4-dicarboxylate_membrane_transporter
C4-dicarboxylate_anaerobic_carrier-like
Protein_of_unknown_function_DUF307
Nitrate_reductase_cytochrome_c-type_subunit_(NapB)
Uncharacterized_lipoprotein
Uncharacterised_protein_family_UPF0208
Protein_of_unknown_function_DUF388,_OB-fold
Protein_of_unknown_function_DUF1090
Tryptophan/tyrosine_permease,_conserved_site
YidE/YbjL_duplication
Protein_of_unknown_function_DUF30/31
Bordetella_pertussis_toxin_A
Mycoplasma_MFS_transporter
Protein_of_unknown_function_DUF687
Chlamydia_15_kDa_cysteine-rich_outer_membrane
Major_outer_membrane_protein,_Chlamydia
Chlamydia_cysteine-rich_outer_membrane_protein_6
Cysteine-rich_outer_membrane_protein_3,_Chlamydia
Uncharacterised_protein_family_UPF0242
Protein_of_unknown_function_DUF1494
Protein_of_unknown_function_DUF1137
Protein_of_unknown_function_DUF1347
Chlamydia_polymorphic_membrane,_middle_domain
Protein_of_unknown_function_DUF1547
Protein_of_unknown_function_DUF1207
Enterotoxin,_bacterial
CblD_like_pilus_biogenesis_initiator
Ferrous_iron_transport_protein,_bacterial
PTS,_lactose/cellobiose-specific_IIB_subunit
Protein_of_unknown_function_DUF441,_transmembrane
Protein_of_unknown_function_DUF1440
PTS,_galactitol-specific_IIC_component
YodA
GlpT_transporter
Ionotropic_glutamate_receptor
Putative_sugar-specific_permease,_SgaT/UlaA
Acid_phosphatase_(Class_B)
Copper_resistance_lipoprotein_NlpE
4-alpha-L-fucosyltransferase
Protein_of_unknown_function_DUF997
Nucleoside:H+_symporter
Glucitol_operon_activator
PTS,_enzyme_II_sorbitol-specific_factor
Sorbitol_phosphotransferase_enzyme_II,_N-terminal
Sorbitol_phosphotransferase_enzyme_II,_C-terminal
Porin,_LamB_type
Uncharacterised_protein_family_UPF0259
Protein_of_unknown_function_DUF1375
GlpM
Cytochrome_b562
Maltose_operon_periplasmic
Secretion_monitor
LPP_motif
Acid_shock
Cell_division_inhibitor_SulA
Protein_of_unknown_function_DUF1434
Protein_of_unknown_function_DUF1418
Protein_of_unknown_function_DUF1471
Intracellular_growth_attenuator_IgaA
Protein_of_unknown_function_DUF1454
PhoQ_Sensor
Protein_of_unknown_function_DUF1131
Protein_of_unknown_function_DUF1198
Protein_of_unknown_function_DUF1435
Phage_shock_protein_G
Protein_of_unknown_function_DUF1422
Protein_of_unknown_function_DUF1481
Protein_of_unknown_function_DUF1144
Conserved_hypothetical_protein,_YtfJ
Cyd_operon_protein_YbgE
WzyE
Antimicrobial_peptide_resistance_and_lipid_A_acylation_PagP
Porin,_general_diffusion_Gram-negative,_conserved_site
Protein_of_unknown_function_DUF1496
Peptidase_S6,_IgA_endopeptidase
HNS-dependent_expression_A
IlvGEDA_operon_leader_peptide
IlvB_leader_peptide
FimH,_mannose-binding
Amino_acid_transporter,_transmembrane
Inhibitor_of_vertebrate_lysozyme
Pili_assembly_chaperone,_conserved_site

Figure 6.17

6.4.1 Functional characterisation of genotypic features overrepresented in microbes successfully thriving in mucosal environments

By studying the IPR domains associated with mucosa-associated microorganisms (as defined in Section 5.1.2) it is possible to define a set of putative functional features that would be required to inhabit a mucosal habitat. In this section, the identified functional features of mucosa-associated protein domains are described. Host mucosal surfaces are naturally covered with mucus. These anatomical barriers are normally enriched with carbohydrates and proteins (see Section 2.2.1). Mucosa surfaces are important interfaces between host internal systems and the external environment. This interface acts as the first protective barrier consisting of a range of physicochemical, and biological defence mechanisms. In order for microorganisms to survive and be able to thrive in such highly-protected environments, microorganisms must be able to access and process nutrients efficiently and at the same time avoid the host defences including the innate and adaptive immune system. The ability to adhere to the specific host cells, tissues or contents in the mucosal environment such as mucus, ECM, saliva, food, faeces, or other host secretions, is also required for the long term survival of microbes in particular niches.

To survive in a specific environment, microorganisms must be able to locate and occupy an optimal niche. Pathogenic strains able to infect hosts via mucosal surfaces are normally equipped with elements facilitating adhesion to the host surfaces, evasion from the immune systems, and invasion to host cells or tissues [Acheson and Luccioli, 2004] [van der Velden *et al.*, 1998] [Peterson, 2002]. In particular, if the microbes originate from different types of habitat before entering the mucosal environment, the invaders need to be able to adapt to survive and thrive successfully in that new environment[Peterson, 2002]. For example, mechanisms for sensing and reacting appropriately to a carbohydrate-riched environment would promote their chance of survival[Houot *et al.*, 2010]. Some pathogens take advantage of the host immune response processes by adapting themselves to tolerate various defensive mechanisms [Cho *et al.*, 2006] [Acheson and Luccioli, 2004].

Carbohydrate transport and metabolic processes

Based on the results of the analysis, domains involved in carbohydrate transport and metabolic processes were identified as expected [Vadeboncoeur and Pelletier, 1997]. More specific cases of an exclusive carbohydrate utilisation are illustrated in the mammalian intestinal commensals. Some mammalian ruminal bacteria such as *B. thetaiotaomicron* and *Ruminococcus flavefaciens* are known to be capable of utilising insoluble structural polysaccharide substrates such as plant cell wall. *B. thetaio-*

taomicron, in particular, can forage on both dietary and host glycans when dietary polysaccharides are not available [Martens *et al.*, 2008] [Flint *et al.*, 2008] [Martens *et al.*, 2009] [Miller *et al.*, 2009].

Carbohydrate metabolism is essential for carbon and energy sources. Many mucosa-associated bacteria are equipped with a specific major carbohydrate transport system, the phosphoenolpyruvate:sugar phosphotransferase system (PTS). Several studies have demonstrated the role of PTS in the control of carbohydrate transport and sugar metabolism in mucosa-associated bacteria including oral, upper respiratory tract, and gastrointestinal tract microbial communities [Vadeboncoeur and Pelletier, 1997] [Bramley and Kornberg, 1987] [Houot *et al.*, 2010]. The PTS is widely distributed across diverse bacterial phyla including the prominent human-specific bacterial phyla such as Actinobcteria, Bacteriodetes, Firmicutes and Proteobacteria.

In the human intestine, a wide range of indigestible dietary plant-associated glycans are conveyed from the upper gut to the large intestine, where the substrates are utilised by the colon commensal communities. SusD (IPR012944; abundance p-value 2.83 $\times 10^{-113}$, co-occurrence p-value 0.005), part of the Bacteriodetes starch utilisation system (Sus) complex, is an outer membrane protein required for the binding of starch to the bacterial cell surface at an early stage of the starch utilisation process [Cho and Salyers, 2001]. Although, the co-occurrence p-value does not pass the cut-off co-occurrence p-value (1 $\times 10^{-4}$) used, the significant abundance p-value suggests the expansion of the SusD protein family in the mucosal microorganisms. The expansion of protein family suggesting the importance of the Sus system in the Bacteriodetes. The SusD domain was mainly abundant in Bacteroidetes phylum, especially among members of normal human gut microbiota. Although, the SusD protein homologs, termed RagB, were identified among the species causing periodontal diseases (e.g. *Porphyromonas gingivalis* and *P. endodontalis*), no sign of protein family expansion was found in the oral pathogenic species in contrast to the gut commensals [Curtis *et al.*, 1999]. The study of RagB in *P. gingivalis* suggested that the protein presented on the cell surface is involved in virulence, although the exact function is still unknown [Nagano *et al.*, 2007] [Curtis *et al.*, 1999].

From the clustering analysis, one cluster of proteins (endo-alpha-mannosidase, hydrolysis of O-glycosyl bond [2]) shared between Bacteroides and *Trichomonas vaginalis* could be involved in mannose metabolism. It is potentially of interest since mannose represents a substantial fraction of mucin carbohydrate moieties of gastric and bronchus mucins [Wagner *et al.*, 1998] (see Section 2.2.1).

[2] `http://www.brenda-enzymes.info/php/result_flat.php4?ecno=3.2.1.101`, accessed 20th August 2010

Peptidase and amino acid transport

Several mucosal microorganisms are known to depend on the exogenous amino acids. For example, the intestinal *Lactobacillus johnsonii* NCC 533 possess a number of duplicated amino acid permeases, peptidase and amino acid transporters to compensate for their lack of genes involved in amino acid biosynthetic pathways [Pridmore *et al.*, 2004]. The organism appears to depend entirely on the host or other local microbes to provide the necessary amino acids for their growth. Microbial secreted and surface proteases may play role in host innate and adaptive immune evasion, extracellular protein hydrolysis and promote adhesion to the mucosal surfaces [Weiser *et al.*, 2003] [Pridmore *et al.*, 2004].

In this study, several amino acid transports, surface peptidases and amino acid permeases were identified as overrepresented in microbes dwelling in mucosal environments. Several of these proteins are distributed across at least two domains of life. For example, a sodium/glutamate symporter (IPR004445; co-occurrence p-value 6.73 $\times 10^{-7}$, abundance p-value 1.28 $\times 10^{-6}$) and tryptophan/tyrosine permease (IPR018227) were found among bacteria and archaea known to be associated with mucosal environments. The sodium/glutamate symporter is a sodium-dependent glutamate uptake. The domain is found among known mucosa-associated archaea (*Methanosarcina spp.*) and bacterial phyla including Actinobacteria, Bacteroidetes, Proteobacteria, Spirochaetes, Firmicutes, Fusobacteria, and Verrucomicrobia.

Tryptophan/tyrosine permease (IPR018227; co-occurrence p-value 7.29 $\times 10^{-9}$, abundance p-value 1.16 $\times 10^{-16}$) is a transporter of aromatic amino acids, mediating cellular import of thryptophan or tyrosine. The domain was found to be encoded by the *tnaB* gene of *Haemophilus influenzae*, a mucosal pathogen [Martin *et al.*, 1998]. The *tnaB* gene is part of the tryptophanase (*tna*) operon, and has been extensively studied in *E. coli*. The *tna* operon encodes genes involved in tryptophanase activity, allowing the thryptophan to be used as a carbon and nitrogen source, resulting in the production of indole, pyruvate and ammonia [Newton and Snell, 1964]. The study by Martin K. *et al.* [Martin *et al.*, 1998] suggested that the *tna* operon may have been acquired by lateral gene transfer. The Thryptophan/tyrosine permease domain was found across bacterial phyla including Actinobacteria, Bacteroidetes, Chlamydiae, Proteobacteria and Firmicutes. The domain was also detected in the non-pathogenic intestinal amoeba, *Entamoeba dispar* SAW760.

Dipeptidase A is a domain that are distributed widely across the three domains of cellular life. Most of the taxa possessing dipeptidase A are known to thrive in a mucosal environment (IPR005322; co-occurrence p-value 1.6 $\times 10^{-5}$, abundance p-value 6.6 $\times 10^{-11}$). The dipeptidase A is a member of

MEROPS Peptidase family C69, clan PB (MER002163)[3]. The domain appears to be abundant among *Lactobacillus spp.* (2-10 copies of the domain). The protein is part of the complex proteolytic system required to obtain essential amino acids [Vesanto *et al.*, 1996].

Signal transduction

Signal transduction is essential to enable microorganisms to sense external stimuli and generate appropriate cellular responses. The role of signalling and rewiring gene expression networks is a rapid adaptive strategy of microorganisms to survive in an environment [Dietrich *et al.*, 2003] [Cases *et al.*, 2003] [Rosenbach *et al.*, 2010].

The PhoQ sensor domain (IPR015014) is a part of the PhoPQ system, a classical two-component signalling system [Cho *et al.*, 2006]. In the case of the animal gut pathogen, *S. typhimurium*, PhoPQ promotes virulence by increasing bacterial tolerance to host antimicrobial peptides and within acidified macrophage phagosomes [Prost and Miller, 2008]. The *S. typhimurium* protein with the PhoQ sensor domain is in the inner membrane and is activated when the bacteria are phagocytosed by the host macrophages or by direct interaction with antimicrobial peptides [Cho *et al.*, 2006]. The sensor responds to the depletion of Mg^{2+} or Ca^{2+}, as well as to acidic conditions in the external environment [Prost and Miller, 2008] [Bearson *et al.*, 1998] [Cho *et al.*, 2006]. The PhoQ domain is restricted to Proteobacteria-Gamma, particularly pathogens e.g. *Shigella spp.*, *Yersinia spp.*, pathogenic *E.coli*, *Pseudomonas spp.*, *Klebsiella spp.*. Based on these data, it is suggested that the PhoQ-containing proteins might have important roles for mucosa-associated intracellular pathogen to adapt rapidly in the host cells during infection.

The ToxR regulatory system in *V. cholera* is another example of known bacterial signal transduction for promoting bacterial survival with in hosts. ToxR coordinates the expression of colonisation and virulence genes in response to specific host signals. The regulatory cascades of the ToxR regulon are not yet well understood [Peterson, 2002]. The ToxR regulon consists of a set of membrane protein sensors for sensing the change of pH, temperature and osmolarity as well as the presence of mucus and bile. However, there are no protein domains related to ToxR presented in the version of InterPro database used in this study. This might be a reason why no ToxR-related domains were detected by this study. The next chapter 7 describes a newly defined domain that is presented in one of the *V. cholera*'s accessory colonisation factors where expression are regulated by the ToxR.

PTS in *Vibrio cholera* is another complex system proven to be important as a signal transduction

[3] `http://merops.sanger.ac.uk/cgi-bin/famsum?family=C69`, accessed 20th August 2010

mechanism used in response to carbohydrate availability, aiding colonisation of the pathogen on gut mucosal surfaces [Houot *et al.*, 2010]. In particular, *V. cholera* senses intestinal mucus as chemotaxins directing the bacteria to move toward the intestinal surfaces and initiating the secretion of proteases capable of degrading mucus [Houot *et al.*, 2010].

Adhesion and colonisation

Motility and attachment to host cells or tissues are important for microbes to move to an optimal environment to initiate and maintain colonisation [Niemann, 2004]. Flagella are structural features of some bacteria that provide the possibility for effective movement as well as adhesion to host [Miron *et al.*, 2001][Bouguenec, 2005]. Colonisation of mucosal surfaces may also be supported by pili or fimbria [Chen *et al.*, 2009][Althouse *et al.*, 2003]. Moreover, some microorganisms aggregate themselves to each other to form a biofilm which enables their attachment to the host or abiotic surfaces [Houot *et al.*, 2010]. The biofilm also contributes to the microbial resistance to host immune system and antibiotics [Anderson and OToole, 2008][Høiby *et al.*, 2010].

Fimbria (or Pili) are another structural feature found on the surface of some bacteria. This appendage facilitates the attachment of the microbe to host surfaces. In some bacteria, fimbriae are required for colonisation to initiate biofilm formations or during infection. Fimbriae are also known as a virulence support factor [Abraham *et al.*, 1998]. An expansion of fimbrial proteins occurred among mucosal-thriving Proteobacteria-Gamma data set (abundance p-value 4.64 $\times 10^{-77}$), suggesting that these structural features play an important role for survival in vertebrate hosts.

The bacterial adhesion domain (IPR008966)[4] is overrepresented among mucosa-thriving microorganisms (co-occurrence p-value 2.1 $\times 10^{-6}$, abundance p-value 3.3 $\times 10^{-97}$). The domain was found in several adhesin proteins. Several protein domains that are regarded as members of this adhesion domain include collagen-binding domain [Symersky *et al.*, 1997], fibrinogen-binding domain [Ponnuraj *et al.*, 2003], fimbrial adhesin lectin domain [Buts *et al.*, 2003], Mannose-binding adhesin [Hung *et al.*, 2002] and PapG adhesin [Dodson *et al.*, 2001].

Microbial adhesins mediating the binding of extracellular pathogens to bind host extracellular matrix proteins are known as MSCRAMMs (microbial surface components recognising adhesive matrix molecules) [Patti and Höök, 1994]. Several bacterial MSCRAMMs play an important role in the development of infection. An example of bacterial adhesins facilitating the adherence of bacteria to vertebrate tissues or mucosa-lined epithelium during pathogenesis including collagen-binding ad-

[4] `http://www.ebi.ac.uk/interpro/IEntry?ac=IPR008966`, accessed 20th August 2010

hesins found in *Staphylococcus aureus* (IPR008456; co-occurrence p-value 0.06, abundance p-value 0.003) [Symersky *et al.*, 1997].

Mannose-binding adhesin (FimH, IPR015243) (co-occurrence p-value 3.9 $\times 10^{-5}$, abundance p-value 7.2 $\times 10^{-7}$) located at the tip of the fimbrillum is an example of bacterial substrate-specific adhesins found on the surface of both commensal and pathogenic Gammaproteobacterial including *E. coli* strains, *Klebsiella pneumoniae* strains, *Proteus mirabilis* and *Shigella spp.*. Mannose is a constituent of gastric and bronchus mucins [Wagner *et al.*, 1998] (see Section 2.2.1).

MUCin-Binding Protein domain (MucBP; PF06458, IPR009459) (abundance p-value 9 $\times 10^{-18}$) was described as a mucus binding component of Lactobacilli. This domain is found across Gram-positive mucosa-thriving bacteria including Lactobacilli, Streptococaceae and *Cryptobacterium curtum*. In particular, MucBP is abundant in gut-specific *Lactobacillus spp.* (abundance p-value 3.2 $\times 10^{-7}$). Repetition of this domain in the same protein suggests increased affinity of adhesins for mucins in the lactic acid bacteria [Boekhorst *et al.*, 2006]. Several proteins containing the MucBP domains were predicted to be involved in binding to mucins or the degradation of complex polysaccharides or mucus-associated glycosylation moieties [Boekhorst *et al.*, 2006].

Lysozyme domain, subfamily 2 (IPR013338)(co-occurrence p-value 5 $\times 10^{-8}$, abundance p-value 1.1 $\times 10^{-10}$), found in Actenobacteria, Firmicutes and Proteobacteria, was shown to hydrolase peptidoglycan and facilitate the formation of flagella rod in *Salmonella typhimurium* [Nambu *et al.*, 1999].

Resistance factors to host defence mechanisms

In order to counter the host defence mechanisms, microbes require stress tolerance factors that provide an advantage in combating against antibacterial compounds, and other stressful conditions in the mucosa environment (e.g. extreme of pH). A variety of commensal and pathogenic Enterobacteria, for example, often encounter acidic stress conditions in the host body [Seputiene *et al.*, 2003]. Proteins encoding acid tolerance genes such as acid shock proteins (ASPs) were found across symbiotic and pathogenic intestinal and urogenital Enterobacteria (see Figure 6.17). ASPs have been shown to increase the ability of microbes to response and survive in an acidic environment [Seputiene *et al.*, 2003]. The acid shock domain (IPR009435) was identified in this study to co-occur and be abundant among Enterobacteria that are able to thrive on mucosal surfaces (co-occurrence p-value 1.4 $\times 10^{-6}$, abundance p-value 4.2 $\times 10^{-7}$).

Conjugated bile salt acid hydrolases (CBAHs), another example of a microbial resistance factor, are enriched among gut microbiota of both bacteria and archaea [Jones *et al.*, 2008]. CBAHs contain a choloylglycine hydrolase domain suggested to be significantly abundant among microbes annotated as mucosal inhabitants (IPR003199; co-occurrence p-value 0.0002, abundance p-value 6.9 $\times 10^{-6}$). The enzyme was demonstrated to enhance the survival of gut-associated microbes *in vitro* by mediating the tolerance of microbes to the host bile acid [Jones *et al.*, 2008]. This domain does not pass the cut-off co-occurrence p-value used (1 $\times 10^{-4}$) for filter out the potential mucosa-associated domains, however, the occurrence p-value is on a border line. This might imply that the method and the cut-off p-value used is quite conservative.

Secreted Immunoglobulin A (sIgA) is an important antibody in mucosal immunity. IgA is the major immunoglobulin found in mucous secretions from saliva glands, mammary glands, gastrointestinal and respiratory epithelium [Acheson and Luccioli, 2004]. IgA endopeptidase (Peptidase S6, IPR000710) is an example of the bacterial host-immune evasion found across Proteobacteria (Beta, Gamma and Epsilon), pathogens and mutualists of human oral, gastrointestinal tract, respiratory and urogenital tracts (see Figure 6.17). The IgA endopeptidase domain is identifed as mucosa-associated in this study, with co-occurrence p-value 1.1 $\times 10^{-5}$, abundance p-value 9.3 $\times 10^{-8}$.

Other resistance factors identified by the analysis conducted in this study included host immune evasion factors such as vertebrate lysozyme inhibitor [Abergel *et al.*, 2007] and antimicrobial peptide (AMP) resistance or lipid A acylation PagP domain (IPR009746; co-occurrence p-value 4.2 $\times 10^{-7}$, abundance p-value 1.1 $\times 10^{-6}$) [Hwang *et al.*, 2002]. Both factors mediate AMP resistance and act as virulent factors. The domains are distributed exclusively among pathogenic mucosa-associated Proteobacteria.

Some mutualists are known to secrete interference factors that increase their resistance to the invasion of some pathogenic strains. This process allows the commensals to sustain their occupancy of the local host environment and also benefits the host with increased protection against pathogens. For example, extracellular serine protease (Esp) from *Staphylococcus epidermidis*, the dominant commensal bacteria in the human nasal cavity, was recently discovered to have an inhibition affect on biofilm formation and nasal colonisation by the pathogenic *S. aureus* [Iwase *et al.*, 2010].

6.4.2 Diversity of survival strategies across mucosa-associated microbial taxonomic groups

Several of the identified extracytoplasmic mucosa-associated protein domains are presented in a specific taxonomic group (see Table 6.4, Appendix H). The restricted taxonomic distribution of these conserved regions might be due to the specific adaptation of microbes in particular conditions. Different groups of mucosa-associated microbes employ different strategies to achieve the functional key features described above. However, the common aim of these actions is to sustain colonisation in the host mucosal environment. The mucosa-associated domains identified in this study are typically distributed amongst a restricted set of taxonomic groups, regardless of whether their members are pathogenic or commensal species (see Figure 6.10). No protein domain was found that was present in all annotated mucosal-associated microbes. These phenotypic characteristics included response to acid shock (IPR009435; discussed earlier) and the presence of a CblD like pilus biogenesis initiator (IPR010888) (see Table 6.4). These two features are specific to mucosa-associated Beta- and Gamma-Proteobacteria. These examples illustrate that each group of the mucosal microbes may have their own mechanisms, shared within a restricted set of taxa, to allow them to survive in mucosal niches. However, it is also important to note that the analysis was performed based on protein domains that were already known or characterised. It is therefore possible that there are some conserved functional regions that have not yet discovered and that these might be shared across a wide range of mucosa-associated microbes from different taxonomic groups.

Mucosa-thriving eukaryotic microbes may utilise different strategies from the bacteria, yet, elements involved in carbohydrate metabolic process are also overrepresented across mucosal eukaryotic microbes. The glycosyl transferase family 35 (GT35; IPR000811) was found across all known mucosa-thriving microbial eukaryotes of both Fungi and Protozoa including Candida, Cryptosporidium, Entamoeba, Trichomonas and Giardia (see Appendix A for the list of eukaryotic taxa). Enzymes from the GT35 family are known to possess glycogen or starch phosphorylase activity [5] (EC 2.4.1.1)[Park *et al.*, 2010]. GT35 is also widely distributed across known mucosa-associated prokaryotes and plant pathogens. Glycosyl transferase, family 31 (GT31) is found across human parasitic protozoa including Trichomonas, Cryptosporidium, Entamoeba. GT31 comprises enzymes with a number of known activities: N-acetyllactosaminide beta-1,3-N-acetylglucosaminyltransferase, beta-1,3-galactosyltransferase, fucose-specific beta-1,3-N-acetylglucosaminyltransferase, and globotriosylceramide beta-1,3-GalNAc transferase [6][Park *et al.*, 2010].

The analysis performed in this study revealed numerous conserved functional features that were either specific to mucosal niches or beneficial to microbes in a carbohydrate-rich or vertebrate host environment. Several known protein signatures of either well-characterised or uncharacterised proteins were shown to be well conserved across microbial species known to associate with mucosal environments either as commensals or pathogens. It would therefore be worthwhile in the future to investigate in more detail the domains of unknown function identified in this study. The character-

[5] http://www.cazy.org/GT35.html, accessed 20th August 2010

[6] http://www.cazy.org/GT31.html, accessed 20th August 2010

isation of these candidate mucosa colonisation domains could lead to a better understanding of the complex mucosa-microbe interactions.

6.4.3 Lateral gene transfer among prokaryotes and microbial eukaryotes sharing mucosal niches

The apparent restricted distribution of several mucosa-associated protein domains across taxonomic classes, particularly prokaryotes and eukaryotes, suggests that lateral gene transfer (LGT) plays an important role for survival among specific mucosa commensals and pathogens [Ragan, 2001] [Andersson, 2009]. The combination of specific functional genotypic features across mucosal microbes from different taxonomic classes were important factors for distinguishing mucosa-associated microbes from other microbes from the same classes. Several of these functional elements were found to be conserved across several mucosa-associated microbes from different and distant taxonomic classes could be explained by LGT. The high-biodiversity and density of microbial communities within mucosal environments, particularly in the human intestinal tract, provides favourable conditions for direct interactions between microbes. An explosive amplification of transposon family among the gut microbial communities also suggests the LGT between the gut microbiota [Kurokawa *et al.*, 2007]. The genetic elements that provide great benefit for the survival of microbes in a particular condition might be strongly selected for and maintained after LGT. For example, an arise of LGT between organisms was suggested in a case of antibiotics resistance genes [Fitzgerald *et al.*, 2001], and genes involved in metabolic enzymes of the substrates enriched in an environment [Guénola *et al.*, 2006] [Hehemann *et al.*, 2010]. The invaders are able to evolve with these essential genetic materials for survival in such a specific condition or environment.

6.4.4 Mucosa-associated protein domains and clusters of unknown function

Many of the domains and clusters identified in this study are unknown function, indicating how much more needs to be discovered about the roles of microbiota in our health and disease. Several of the unknown function protein domains are presented in a restricted taxonomic groups might be due to the specific adaptation in a particular condition.

Among the unknown function of protein clusters, several of them do not have any defined conserved regions. They represent candidates of important function to thrive in mucosal environment, therefore might be interesting to identify their function. Further detailed bioinformatics analysis, such as network-based prediction of protein function [James *et al.*, 2009][Sharan *et al.*, 2007], can performed to generate hypotheses about their functions. These *in silico* analyses can be combined with web-lab experiments including gene expression pattern in different conditions [Martens *et al.*, 2008], transposon mutagenesis for connecting phenotype to gene [Goodman *et al.*, 2009].

6.4.5 Future perspective

The mucosal domain analysis results represents a proof of principle that association analysis can be used to characterise important molecular functions of microorganisms for their survival in a particular habitat. Several protein domains were identified that are already known to be essential for the microbes during their interaction with host mucosal environments. Moreover, some of the results from this study agree with a recent publication reporting the discovery of conserved regions across a newly sequenced random shotgun human gut metagenomic data set [Ellrott *et al.*, 2010].

The results presented in this chapter could be complemented by more fined-grain analysis in which mutualistic and pathogenic taxa are contrasted [Rasko *et al.*, 2008]. Such analyses might provide additional insights into the molecular basis of microbial factors that are beneficial to our health as well as involved in the pathogenesis, respectively. It is noticeable that so far, each particular group of bacterial commensals provide different benefit to hosts. Of the dominant nasal cavity commensals, *S. epidermidis* has been shown to release an interference factor that has an inhibitory role on the nasal colonisation and biofilm formation of pathogenic *S. aureus* [Iwase *et al.*, 2010]. Similar scenarios can be seen in the pathogenic strains which have adverse effects on the host body through different strategies and mechanisms. Several known gut Gamma-proteobacterial pathogens are equipped with enterotoxins [Chapman *et al.*, 2006]. Therefore, proteins that are unique or exclusively presented within a species or strains are worth investigating.

Beyond providing a global view of the molecular functions of mucosal microorganisms, particularly for the gut microbiome, an extensive list of the mucosa-associated protein domains and families established by this work enables future studies of both laboratory and computational experiments. These studies will lead to a better understanding of the vertebrate host-microbe interactions. For example, the clusters of unknown protein domains containing mucosa-associated proteins identified in this study can be used as a guide to narrow down a list of candidate of uncharacterised proteins that are potentially involved in the survival of microbes in a mucosal environment.

The approach used in this study is able to identify features overrepresented in a broad range of annotated mucosa-thriving microbes, as well as restricted groups of microbes. The approach discovered protein domains unique to some phyla that are known to cause disease via mucosal surfaces, such as

Chlamydia and Mycoplasma. Several virulence factors specific to those exclusive bacterial groups were shown in the list of mucosa-associated protein features (see Table 6.4 and Appendix H). However, these features are not necessary associated with the mucosa-thriving ability of the microbes. For example, some mucosal pathogens might have multiple hosts or interact with several other environments such as *Yersinia spp*. and *Chalmydia spp* [Pallen and Wren, 2007]. Domains that are shared among several mucosal phyla are more likely to play an important role in the long-term survival of a microbe in the mucosal environment.

The taxa included for the clustering analysis in this study was restricted to 75 known mucosa-thriving taxa. It would be interesting to perform the clustering analysis with additional genomes in order to expand the views of the distribution of the extracytoplasmic proteomes among mucosal microbes. Increasing the number of taxa sampling, the more meaningful comparative genomics analysis results can be obtained [Tatusov *et al.*, 2003].

Finally, several protein domains and proteins of unknown function were identified as mucosa-associated elements. These findings reinforce the notion that there are still many more important functional features to be discovered among mucosal microorganisms. Furthermore, several metagenomics studies of the human gut microbiome have revealed large fraction of functionally uncharacterised proteins [Ellrott *et al.*, 2010][Kurokawa *et al.*, 2007]. These findings would contribute to the prioritisation of future more detailed bioinformatics analyses and functional characterisation through experimental works.

6.5 Conclusions

The comparative analysis performed in this study has revealed conserved functional elements that are potentially important for microorganisms to survive in vertebrate mucosal environments. The study highlighted some principles of the mucosa-microbe interactions from the perspective of microbial extracytoplasmic proteomes. The approach not only identifies known traits that are important for the survival of microbes in the mucosal environments, it also reveals previously unidentified, ‘novel’ conserved protein regions, that are potentially specific to mucosal microorganisms. In order to initiate and sustain successful colonisation of the highly defended host mucosal niches, microbes must possess a combination of features performing a variety of biological processes and molecular functions. These features include the ability to metabolise and transport carbohydrates and proteins which are the typical substrates found in mucosal environments. Signal transduction in response to environmental cues, such as host immune responses and changes in nutrients concentration, is also a feature essential for the mucosal microbes. The highly-defended host immune system presents various environmental pressures for both commensal and pathogenic mucosal microbial communities. Therefore, resistance factors enabling the microbes to survive these external stresses are also required. Another feature facilitating the successful habitation of mucosal environments is the ability of the microbes to move and colonise the host surfaces. Some mucosal microbes are found to be equipped with sugar- or mucin- binding adhesins. Several of these microbes have flagella or fimbria which facilitate their movement and attachment to the host surfaces. Gene duplication and lateral gene transfer are important evolutionary events driving the evolution of mucosa-associated microorganisms. Indeed, several mucosa-associated protein families have undergone dramatically gene duplication events with over 50 paralogs encoded in a single genome. Several key features involved in specific metabolisms appear in both prokaryotes and distantly related microbial eukaryotes, suggesting the genes were acquired via LGT. Several of these features are likely to enable microbes to become specialised to mucosal environments and are therefore necessary, but may not sufficient, for a survival of a microbe in a specific mucosal ecological niche.

Chapter 7

A novel zinc-metalloprotease-like domain in host-associated microbes and a new functional context for carbohydrate binding modules

7.1 Introduction

Trichomonas vaginalis is a mucosa-associated microbial eukaryote, which causes the most common, non-viral, sexually transmitted infection (STI) [Schwebke and Burgess, 2004][Johnston and Mabey, 2008]. The completed draft genome sequence of *T. vaginalis* G3 was recently published [Carlton *et al.*, 2007], and initial annotations of the *T. vaginalis* genome have reported a set of candidate surface proteins potentially involved in host-pathogen interactions that are similar to sequences with known microbial surface proteins [Carlton *et al.*, 2007][Hirt *et al.*, 2007]. One family of *T. vaginalis* candidate surface proteins showed significant sequence similarity to a *Entameoba histolytica* immuno-dominant protein in BLAST searches [Hirt *et al.*, 2007]. The immuno-dominant variable surface antigen identified in *E. histolytica*, a parasite of the human lower digestive tract [Edman *et al.*, 1990], was experimentally shown to be recognised by more than 70% of immune sera from patients with an amoebic abscess [Edman *et al.*, 1990]. In *E. histolytica*, the immuno-dominant protein was hypothesised to act as a parasite surface receptor for the phagocytosis of human apoptotic cells, and proteomics analysis of the parasite phagosomes indicating the protein was located in the phagosomes during the invasion process [Marion and Guillén, 2006]. However, the function of the *E. histolytica* immuno-dominant surface protein is currently unknown. The presence of candidate surface proteins with sequence features shared between two mucosal parasite members of distant major eukaryotic lineages [Adl *et al.*, 2005] raised the question of whether these related sequences are shared among organisms that thrive on animal hosts and whether these proteins could play an important role in host-microbes interactions.

In silico characterisation of proteins related to immuno-dominant proteins from the parasitic protozoa revealed a novel mucosa-associated protein domain, we named ‘M60-like’. The name ‘M60-like’ was used due to the finding that the new domain is distantly related to a characterised protease family, M60-metallopeptidase enhancin (Pfam:PF03272). The M60-like protein domain was detected among microorganisms inhabiting mucosa-lined niches, as well as animal hosts possessing mucosal epithelial layers. Bioinformatics analyses of the M60-like domain identified a conserved motif with a potential catalytic function relating to a gluzincins metalloprotease. Extracellular or cell-surface targeting signals were detected in microbial proteins carrying M60-like domains, indicating that the proteins are either secreted or expressed on the cell surface. Mucosa-related Carbohydrate-Binding Module (CBM), CBM32 and CBM5_12, were also identified on several M60-like-containing proteins encoded by known mucosal inhabitants or pathogens. The co-occurrence of the CBMs and M60-like domain reveals a new functional context for CBMs, which have previously been associated with carbohydrate processing enzymes, but not proteases.

A M60-like HMM profile was constructed and deposited in the Pfam database with accession PF13402[1]. This chapter describes the novel M60 protein domain, which may be of interest to future studies addressing the context of host-microbe interactions or mucosal colonisation, as well as targeting molecules for the conserved gluzincin metallopeptidase.

7.2 Methods

7.2.1 Sequence similarity search and HMM profile generation

To identify proteins related to the *T. vaginalis* Immuno-dominant variable surface antigen-like proteins, a homolog of the *T. vaginalis* protein (NCBI accession: XP_001313628.1; GI|123449825; UniProt Accession: A2F335) was used as a query to perform a PSI-BLAST search against the NCBI RefSeq database (search date: January 20th, 2010). Only the first 500 amino acids were found to be conserved across a broad range of taxa. This conserved region was then used to perform the

[1] `https://pfamsvn.sanger.ac.uk/svn/pfam/trunk/Data/Families/PF13402/`, accessed 15th December 2010

PSI-BLAST search. A multiple sequence alignment of all the PSI-BLAST (one iteration) hit protein sequences with an e-value cut-off 1×10^{-4} were retrieved from the BLAST server. The segment of the aligned conserved sequences corresponding to positions 131-431 of the *T. vaginalis* query sequence was identified as the most conserved region across the alignment. Sequences annotated with M60-enhancin domains (PF03272) were removed from the alignment to ensure that there was no overlap between the new domain and enhancin protein family. Sequences with identity level $\geq 80\%$ were considered as highly-related and the shortest one was removed from the alignment. After the sequences were removed, 68 sequences remained in the alignment [Bateman, 2010, pers. comm.]. HMMER3[2] [Eddy, 1998] was then employed to generate and calibrate a new HMM profile, named M60-like, from the alignment of the conserved region.

In order to identify CBM5_12 and CBM32 domains on protein sequences containing M60-like domains, the HMM profiles representing CBM5_12 (SSF51055) and CBM32 were derived from SUPERFAMILY database [Wilson *et al.*, 2009]. For the CBM32 profile, five HMM models (0036212, 0036298, 0043558, 0043559, 0047789) from the SUPERFAMILY were used. Each model was annotated as 'Discoidin domain (FA58C, coagulation factor 5/8 C-terminal domain)'. These five models are part of the 'Galactose-binding domain-like superfamily' (SSF49785). For the CBM5_12 HMM profile, three SUPERFAMILY HMM models representing the SSF51055 (0035067, 0036915, 0036705) were used.

7.2.2 Detection of functional protein regions in proteins containing M60-like domains

Phobius and TMHMM 2.0 were employed to detect extracellular-targeting N-terminal signal peptide and alpha-helix transmembrane regions. LipoP 1.0 was used to predict an N-terminal signal peptidase II cleavage site of a lipoprotein candidate. InterProScan version 4.4 was used to search for other characterised protein domains and motifs. The default parameters were used for every tool.

7.2.3 Protein profile HMM searches

HMMER3 was used to search M60-like HMM profile against proteins in RefSeq database (data obtained on 21th January 2010 from ftp://ftp.ncbi.nih.gov/blast/db, containing 9,662,677 protein sequences). The HMM profiles of M60-like, CBM32 and CBM5_12 were also searched over an annotated protease library retrieved from the MEROPS database (file obtained 2nd May 2010, containing

[2] `http://hmmer.wustl.edu/`, accessed 1st December 2010

177,390 sequences). The 'hmmsearch' command was used to search the profiles against both the RefSeq and MEROPS protein sequences. An e-value of $< 1 \times 10^{-5}$ was used as an inclusion criteria.

7.2.4 Protein profile-profile searches

To perform HMM-HMM profile comparisons between the M60-like profile and other known HMM profiles, HHPred server (see Background section 2.8.3) running with HHSearch version 1.6.0.0 was used to search InterPro database version 16.2. The 'global alignment' option was used to search for potential homologous protein domains.

7.2.5 Associating the M60-like domain to microbial mucosal-related lifestyle

To investigate the significance of the association between the presence of an M60-like domain (genotype) and mucosal-related lifestyle (phenotype) of microorganisms, the probability of the co-occurrence between the protein domain and the phenotypic feature was calculated using the hypergeometric distribution function (see section 2.9.1). The hypergeometric test was used to assess the probability of finding the M60-like protein domain in the annotated mucosa-associated microbes compared to the number of other habitat-classified organisms with the protein domain.

The number of organisms that are known to be associated with animal hosts or to be more specific, vertebrate mucosa surfaces, can be summarised (Figure 7.1, more details in Appendix I). These numbers were summarised considering information about the habitat or isolation source of microorganisms according to the GOLD database (derived 22nd October 2009),

The equation for the hypergeometric distribution is:

$$p(i \geq m \mid N, M, n) = \sum_{i=m}^{n} \frac{\binom{M}{i}\binom{M-n}{n-i}}{\binom{N}{n}}$$

Of the total number of microorganisms with completed genome sequences in the RefSeq database, 455 (N) have habitat information that can be used to determine whether an organism is able to thrive on or penetrate through vertebrate mucosa surfaces. The number of these microorganism with an M60-like domain annotated was 62 (n). The number of microorganisms known to thrive on or infect host through mucosal surfaces was 197 (M). Of these 197 taxa, 45 (m) taxa possess at least one M60-like domain. As a result, the probability (p-value) of observing the association of the M60-like domain and the ability of microbe to thrive on mucosal surface is 3.7×10^{-9} (see Figure 7.1). This genotype-phenotype association may have either a positive or negative direction, which represent the presence or absence of the M60-like domain facilitating the mucosal lifestyle of microbes. To determine the type of this association, the mean value (μ) of the hypergeometric distribution was calculated (see Section 2.9.1). The mean value can be calculated by:

$$\mu = n * M/N$$

Where, m > (μ) shows a positive association and m < (μ) illustrates a negative association. In our case, (μ) is 26.8 (62*197/455) which is less than 45 (m). It can be concluded that the presence of an M60-like domain can be associated with the mucosal phenotype of microbes, and this association is statistically significant.

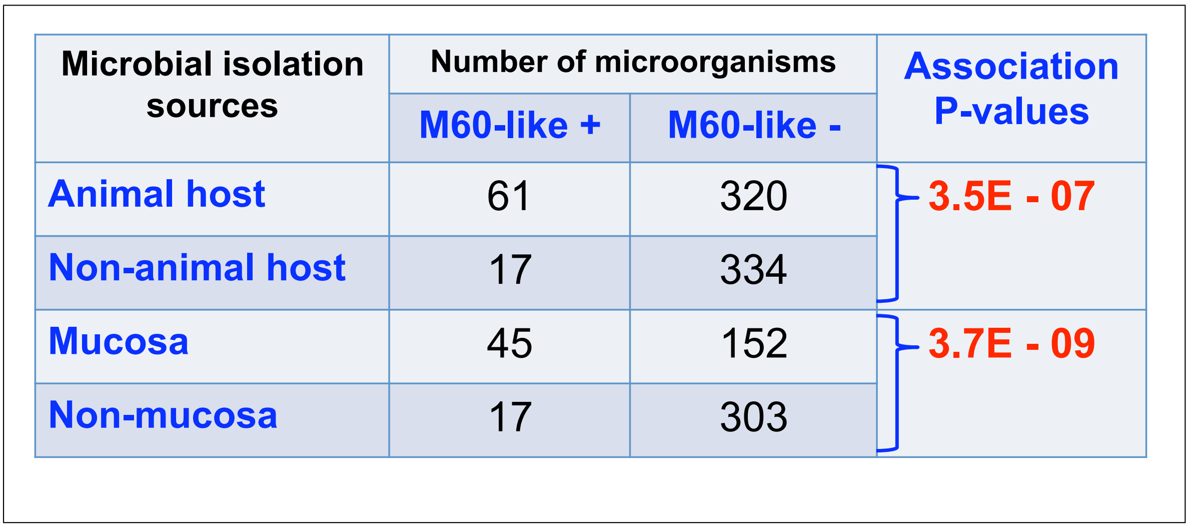
The same approach was also applied to find whether there is a positive association between the M60-like domain and animal-host associated microorganisms. Given the number of organisms with a complete genome sequence for which habitat or isolation source information is available (N), N is 654. The number of these microorganisms with M60-like domains (n) was 78. The total number of microorganisms known to associate with animals (M) was 320. The number of microorganisms that have both phenotype and genotype was 61 (m). As a result, p-value of the association of M60-like and animal-associated microbes was 3.5×10^{-7} (see Figure 7.1). In this case, μ is 38.2 (78*320/654) which is also less than 61 (m). This result suggests a significant positive association between the presence of the M60-like domain in the animal-associated microorganisms.

7.3 Results

7.3.1 Identification of the M60-like protein domain and construction of HMM profile

To identify a potential conserved region of the surface immuno-dominant proteins, a set of proteins from *T. vaginalis* [Carlton *et al.*, 2007][Hirt *et al.*, 2007], that share sequence features with the protein from *Entamobea histolytica* [Edman *et al.*, 1990] on the basis of BLASTP hits, was used as a query to perform BLASTP search. The following most significant hits included proteins from bacteria known to be able to thrive on mammalian mucosal surfaces including: *Mycoplasma penetrans* (a Mollicute) a human mucosa pathogen that can infect the urogenital and respiratory tracts; [Sasaki *et al.*, 2002] and *Clostridium perfringens* (a Firmicute) that can infect the digestive tract of various mammals [Brynestad and Granum, 2002] as well as mammalian sequences.

Performing one-iteration PSI-BLAST search with an e-value cut-off of less than 1×10^{-4}, 552 hits to protein sequences from 333 different species and strains. The hit list was characterised by a highly



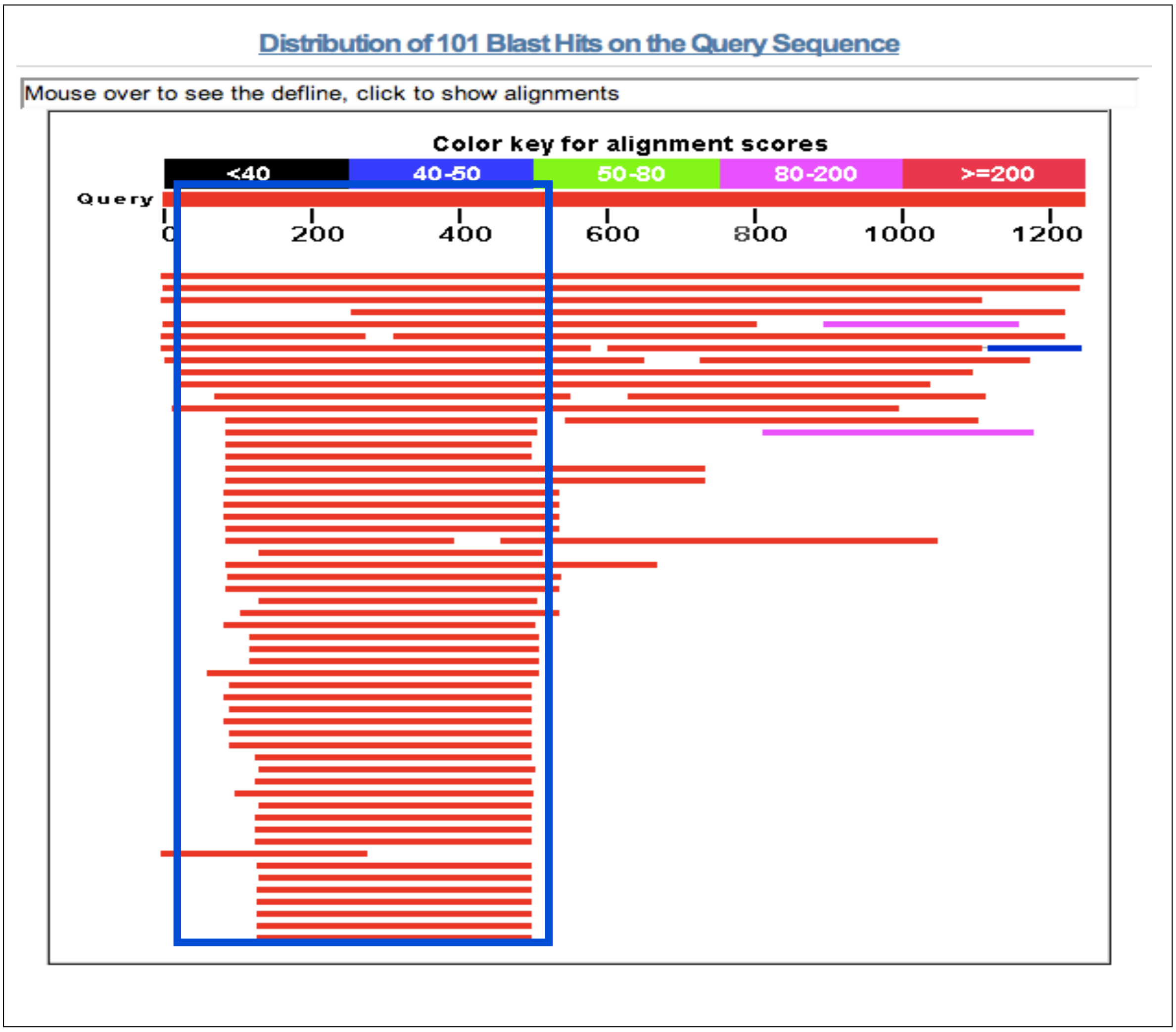
Microbial isolation sources	Number of microorganisms		Association P-values
	M60-like +	M60-like -	
Animal host	61	320	3.5E - 07
Non-animal host	17	334	
Mucosa	45	152	3.7E - 09
Non-mucosa	17	303	

Figure 7.1: Significance scores of the association of the M60-like domain and host-associated microbes. The M60-like domain is significantly associated with microbes living on animal hosts, in particular vertebrate mucosa surfaces. The association values were calculated using hypergeometric test. The p-value produced from the test represents the probability of finding the M60-like domain in the test set in relation to the reference set. To assess whether there is a significant association between the domain and mucosa-associated taxa, the number of mucosa-associated taxa was used as a test set compared to the number of taxa of that are either non-mucosa or mucosa associated. The association between the domain and host-associated microorganisms was also evaluated. The number of animal host-associated taxa was used as a test set in relation to the number of all taxa with known habitats as a reference set.

patchy taxonomic distribution containing a broad mix of eukaryotes, bacteria and baculoviruses. The largest hit lists for a given taxon are from *T. vaginalis* and *Bacteroides caccae*, with 26 and 16 sequences, respectively.

An alignment of the PSI-BLAST hit results showed that residues at position 100-500 at the N-terminus of the *T. vaginalis* query protein sequence (RefSeq accession: XP_001313628, 1247 residues) co-aligned with sub-regions of related sequences from other mucosa-associated organisms (see Figure 7.2). However, no known functional motifs or domains were detected in the corresponding segment when scanning the query sequence against an InterPro integrated database of protein domains and functional sites. The absence of recognised features on the broadly conserved regions suggested the discovery of a potentially new protein domain. A more specific HMM profile of the M60-like protein domain was generated based on a multiple sequence alignment of the conserved region (length of 198 amino acids) of the non-redundant sequences from similar proteins retrieved from one-iteration of PSI-BLAST hit results (Appendix K).

To investigate the features of the conserved sequence region, a multiple sequence alignment was generated with the sequences from the PSI-BLAST hit list that maximised site homology and removed highly similar and partial sequences (see Section 7.2.1 for details) over the conserved segments. This process resulted in an alignment composed of 387 columns across 68 sequences that was used to



Distribution of 101 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments

Figure 7.2: The M60-like conserved region from the BLASTP search results. The *Trichomonas* M60-like protein (XP_001313628.1; GI|123449825) was used as the query for the PSI-BLAST search (1 iteration). The blue box highlights a well-conserved region across proteins from other host-associated microorganisms, in particular, mucosa-associated microbes. This conserved region had not yet been previously characterised.

generate a new profile of a protein domain. The new protein domain was then deposited in Pfam with the accession number PF13402[3].

7.3.2 The M60-like domain is related to the M60-enhancin Zn-metalloproteases

HMMER was used to search with the newly generated M60-like HMM profile the RefSeq protein database (retrieved date: 20th January 2010). This search identified 523 significant hits (e-values < 1×10^{-5}) derived from 322 taxa. Taxa annotated with the M60-like domains included members of seven major bacterial taxa (124 Firmicutes, 144 Proteobacteria, 18 Bacteroidetes, 3 Verrucobacteria, 2 Actinobacteria, 1 Planctomycetes, 1 Tenericutes), and eukaryotic taxa (14 Metazoa, 3 Fungi, 2 Amoebozoa, 2 Apicomplexa, 1 Parabasala and 1 Choanozoa) (see Appendix J).

The vast majority of identified proteins containing M60-like domains, 489 entries (93.5%), possessed the minimal HEXXH zincin motif that was aligned to each other in a global alignment (Appendix K). This motif is characteristic of a broad range of functionally characterised Zn-metallopeptidases with the two histidine residues being ligands of a catalytic Zn^{++} atom and a glutamate representing the single catalytic residue [Jongeneel *et al.*, 1989][Bode *et al.*, 1993][Gomis-Rüth, 2003]. An additional conserved glutamic acid residue was also aligned across the related sequences and was found within 28 residues C-terminally to the zincin motif defining the pattern HEXXHX...E (Appendix K). The conserved consensus HEXXH(8,28)E motif is suggestive of a gluzincin-like family of Zn-metallopeptidases, where the second conserved glutamate potentially acts as a third protease zinc ligand [Hooper, 1994].

The presence of gluzincin sequence features in the M60-like domain prompted us to search the MEROPS database with the newly generated M60-like profile [Rawlings *et al.*, 2008] to investigate the presence of the domain in known proteases. Using HMMER to perform the search with a cut-off e-value < 1×10^{-5}, 38 positive MEROPS entries were found for the M60-like domains. Twenty-one are members of the family M60 unassigned peptidase (enhancin), while the remaining 17 entries are annotated as enhancin-like peptidases (see Table 7.1 and Appendix L). The predicted regions of M60-like domains on the proteases identified by HMMER partially overlap with the regions responsible for peptidase activity. The 38 sequences from the MEROPS database with a positive HMMER result were then analysed with InterProScan to determine whether an M60-enhancin (PF03272) was also present. Only three of MEROPS sequences did not hit the M60-enhancin domain; these were the MEROPS hits with the top three HMMER scores (see Table 7.1).

[3] `https://pfamsvn.sanger.ac.uk/svn/pfam/trunk/Data/Families/PF13402/`, accessed 15th December 2010

Table 7.1: MEROPS proteases encoding M60-like domains. The table contains MEROPS identifiers with their descriptions. Scores and e-values of the three most significant hits resulting from a HMMER search of the M60-like profile are shown (See Appendix L for a complete hit list). Carbohydrate-binding modules (CBMs) are also listed (if present) with their predicted locations.

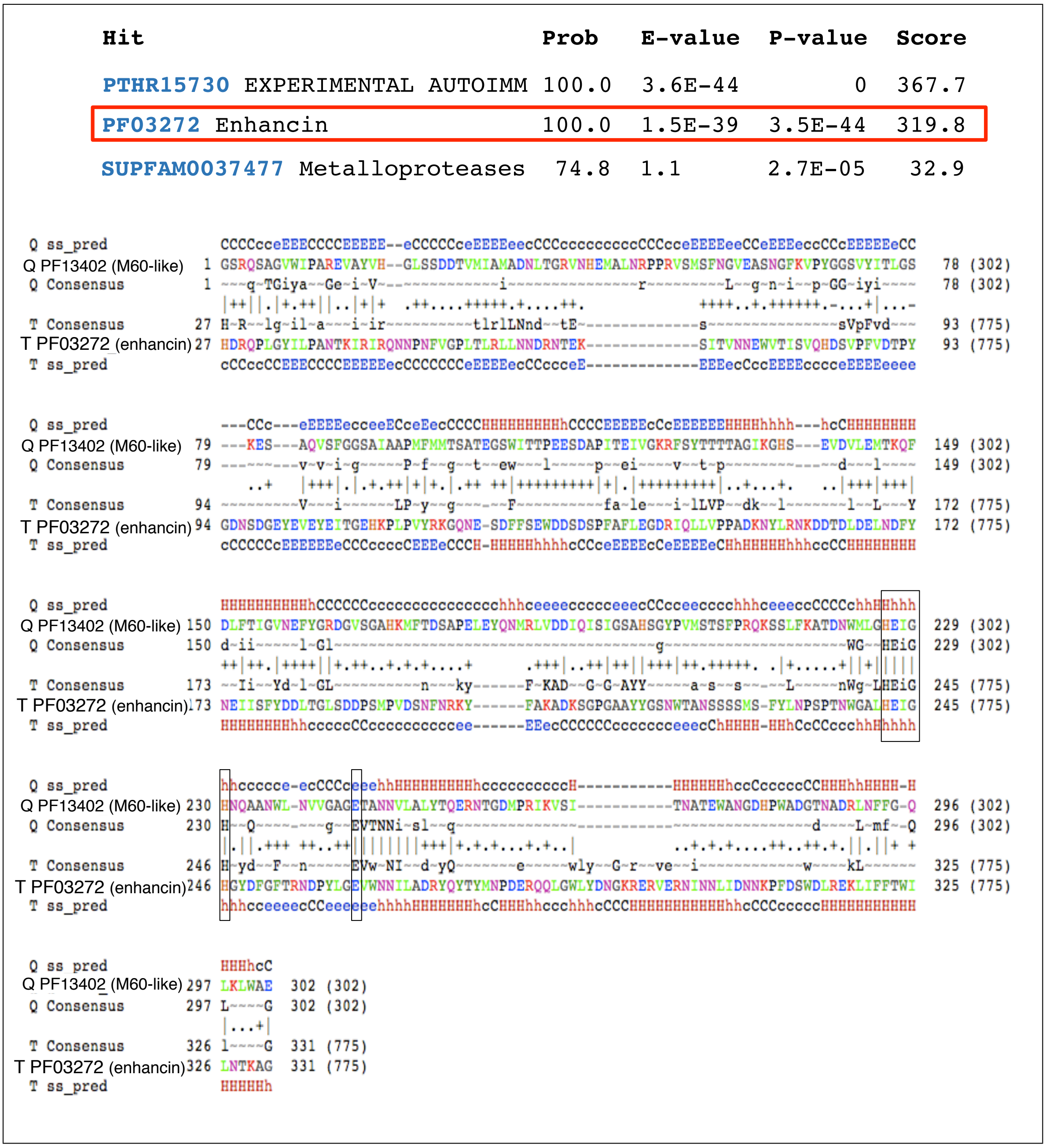
Merops ID	Description	Organism	M60-like start-end	score	e-value	CBM
MER151941	family M60 unassigned peptidases (peptidase unit: 223-427) from GB:ACK98449	*Bacillus cereus*	82-377	310.5	1.1×10^{-91}	none
MER150257	family M60 unassigned peptidases (peptidase unit: 209-420) from GB:ACK63685	*Bacillus cereus*	75-370	292.2	2.5×10^{-86}	CBM32 449-591
MER111749	family M60 unassigned peptidases (peptidase unit: 250-458) from GB:ACD04464	*Akkermansia muciniphila*	100-403	229.5	5.1×10^{-67}	none

The presence of the HEXXHX(8,28)E pattern in 92% (482/523) of the M60-like positive RefSeq sequences (including the *E. histolytica* entry), was found in a range of proteins annotated as M60-enhancin from the PSI-BLAST search. Together with the 38 enhancin-like proteins in the MEROPS database which possessed the M60-like profile, the evidence strongly suggests that M60-like positive proteins are proteases.

To further investigate this possibility, HMM-HMM profile comparison of the M60-like and M60-enhancin was performed using HHPred [Soding, 2005]. Significant hits for the M60-like domain were recovered for several profiles by searching all databases available on the HHPred server. The first hit corresponds to a domain with no known assigned function. The second and third hits correspond to M60-enhancin proteases where the aligned positions between the M60-like and the M60-enhancin profiles included the motif HEXXHX(8,28)E (Figure 7.3). Taken together, these different considerations also support the hypothesis that the M60-like domain corresponds to a newly identified Zn-metallopeptidase family that is distantly related to the M60-enhancin family.

7.3.3 The M60-like protein domain is widely distributed across host-associated organisms

The 523 protein sequences containing the M60-like profile were derived from 322 taxa across bacteria and eukaryotes. The majority of taxa encoding proteins with M60-like domains are microorganisms known to be either commensals, mutualists or pathogens of animal hosts including vertebrate mucosa or invertebrate digestive tracts (chitin-containing) (see Appendix J). Some species are able to thrive on both insect and mammalian hosts, for example, *Yersinia enterocolitica* [Heermann and Fuchs, 2008]. Indeed, a highly significant positive association between the M60-like



```
Hit                                     Prob   E-value   P-value   Score

PTHR15730 EXPERIMENTAL AUTOIMM         100.0   3.6E-44         0   367.7
PF03272 Enhancin                       100.0   1.5E-39   3.5E-44   319.8
SUPFAM0037477 Metalloproteases          74.8   1.1       2.7E-05    32.9


 Q ss_pred                   CCCCcceEEECCCCEEEEE--eCCCCCceEEEEeecCCCcccccccccCCCcceEEEEeeCCeEEEeccCCcEEEEEeCC
Q PF13402 (M60-like)     1  GSRQSAGVWIPAREVAYVH--GLSSDDTVMIAMADNLTGRVNHEMALNRPPRVSMSFNGVEASNGFKVPYGGSVYITLGS   78 (302)
 Q Consensus              1  ~~~q~TGiya~~Ge~i~V~-~~~~~~~~~~~~i~~~~~~~~~~~~~~~~r~~~~~~~~~~~L~~g~n~i~~p~GG~iyi~~~~   78 (302)
                             |++||.|+.++||..|+|+  .++....+++++.+....++.             ++++..+.++++++.-...+|...-
 T Consensus             27  H~R~~lg~il~a~~~i~ir~~~~~~~~~~tlrlLNnd~~tE~-------------s~~~~~~~~~~~~~~~sVpFvd~~~   93 (775)
T PF03272 (enhancin)    27  HDRQPLGYILPANTKIRIRQNNPNFVGPLTLRLLNNDRNTEK-------------SITVNNEWVTISVQHDSVPFVDTPY   93 (775)
 T ss_pred                   cCCccCCEEECCCCEEEEEecCCCCCCeEEEEecCCcceE-------------EEEecCccEEEEccccEEEEeeee


 Q ss_pred                   ---CCc---eEEEEecceeECceEecCCCCHHHHHHHHHhCCCCEEEEEcCcEEEEEEHHHHhhhh---hcCHHHHHHHH
Q PF13402 (M60-like)    79  ---KES---AQVSFGGSAIAAPMFMMTSATEGSWITTPEESDAPITEIVGKRFSYTTTTAGIKGHS---EVDVLEMTKQF  149 (302)
 Q Consensus             79  ---~~~~~~v~v~i~g~~~~~P~f~~g~~t~~ew~~~l~~~~~p~~ei~~~~v~~t~p~~~~~~~~~~~~~d~~~l~~~~  149 (302)
                                ..+    |+++|.|.+.++|+|+.|.++ ++|+++++++++|+|.|+++++++++|..+...+.   ..|+++|+++|
 T Consensus             94  ~~~~~~~~~V~~~i~~~~~~LP~y~~g~~~~~~F~~~~~~~~~~fa~le~~~i~lLVP~~dk~~l~~~~~~~~l~~L~~~Y  172 (775)
T PF03272 (enhancin)    94  GDNSDGEYEVEYEITGEHKPLPVYRKGQNE-SDFFSEWDDSDSPFAFLEGDRIQLLVPPADKNYLRNKDDTDLDELNDFY  172 (775)
 T ss_pred                   cCCCCCcEEEEEEeCCCccccCEEEeCCCH-HHHHHhhhhcCCceEEEEcCeEEEEeCHhHHHHHhhhccCCHHHHHHHH


 Q ss_pred                   HHHHHHHHHHHhCCCCCCcccccccccccccccchhhceeeecccccceeecCCcceeccccchhhceeeccCCCCchhHHhhh
Q PF13402 (M60-like)   150  DLFTIGVNEFYGRDGVSGAHKMFTDSAPELEYQNMRLVDDIQISIGSAHSGYPVMSTSFPRQKSSLFKATDNWMLGHEIG  229 (302)
 Q Consensus            150  d~ii~~~~~l~Gl~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~g~~~~~~~~~~~~~~~~~~~~~~~~~~~WG~~HEiG  229 (302)
                             ++|++.|++++||+.++..+.+.+....+      .+++|..++|++||+++|++.+++++. .|+.....+||+|||||
 T Consensus            173  ~~Ii~~Yd~l~GL~~~~~~~~~~n~~~ky------F~KAD~~G~G~AYY~~~~~a~s~~s~~~~L~~~~~nWg~LHEiG  245 (775)
T PF03272 (enhancin)   173  NEIISFYDDLTGLSDDPSMPVDSNFNRKY------FAKADKSGPGAAYYGSNWTANSSSSMS-FYLNPSPTNWGALHEIG  245 (775)
 T ss_pred                   HHHHHHHHhhcccccCCcccccccccee------EEecCCCCCCcccccccceeecChHHHH-HHhCcCCccchhHhhhh


 Q ss_pred                   hhcccccce-ecCCCceeehhHHHHHHHHHhccccccccccH-----------HHHHHHhccCccccCCHHHhhHHHH-H
Q PF13402 (M60-like)   230  HNQAANWL-NVVGAGETANNVLALYTQERNTGDMPRIKVSI-----------TNATEWANGDHPWADGTNADRLNFFG-Q  296 (302)
 Q Consensus            230  H~~Q~~~~~~~~g~~EVTNNi~sl~~q~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~d~~~~L~mf~~Q  296 (302)
                             |.||.+++ ++..++||||||||+++|+.+.+...+.+..|           ..+.+.+....++..++.+.||.||+ +
 T Consensus            246  H~yd~~F~~n~~~~~EVw~NI~~d~yQ~~~~~~~e~~~~~wly~~G~r~~ve~~i~~~~~~~~~~~~w~~~~kL~~~~~~  325 (775)
T PF03272 (enhancin)   246  HGYDFGFTRNDPYLGEVWNNILADRYQYTYMNPDERQQLGWLYDNGKRERVERNINNLIDNNKPFDSWDLREKLIFFTWI  325 (775)
 T ss_pred                   hhhcceeeeeCCeeeeeehhhhHHHHHHHHhcCHHHhhccchhhcCCCHHHHHHHHHHHHhhcCCCcccccHHHHHHHHHHH


 Q ss_pred                   HHHhcC
Q PF13402 (M60-like)   297  LKLWAE   302 (302)
 Q Consensus            297  L~~~~G   302 (302)
                             |...+|
 T Consensus            326  l~~~~G   331 (775)
T PF03272 (enhancin)   326  LNTKAG   331 (775)
 T ss_pred                   HHHHHh
```

Figure 7.3: Profile-profile alignments of the M60-like domain and M60-enhancin profile. The profile alignment was derived from a HHpred search against all protein signature databases, with an M60-like domain alignment used to generate the HMM profile. The top three hit list are shown at the top of the figure. The alignment of M60-like and M60-enhancin HMM profiles are shown. The HEXXH...E catalytic motif were highlighted in square boxes. This motif is well conserved across the two profiles suggesting catalytic function of both protein profiles. The profile alignment consists of 'SS_pred' lines representing secondary sequence structures predicted by PSIPRED, as well as 'consensus' lines showing the consensus sequences of the M60-like domains and the corresponding hit domains. Amino acid residues are marked in capital letters when they occur over 60% of the corresponding alignment, and in lower case when they are presented greater than 40% of the alignment. Tilda indicates unconserved columns. The line in between the two consensus sequences shows the match quality and is defined as follows: '=' very bad match, '-' bad, '.' neutral, '+' good match and '|' very good match.

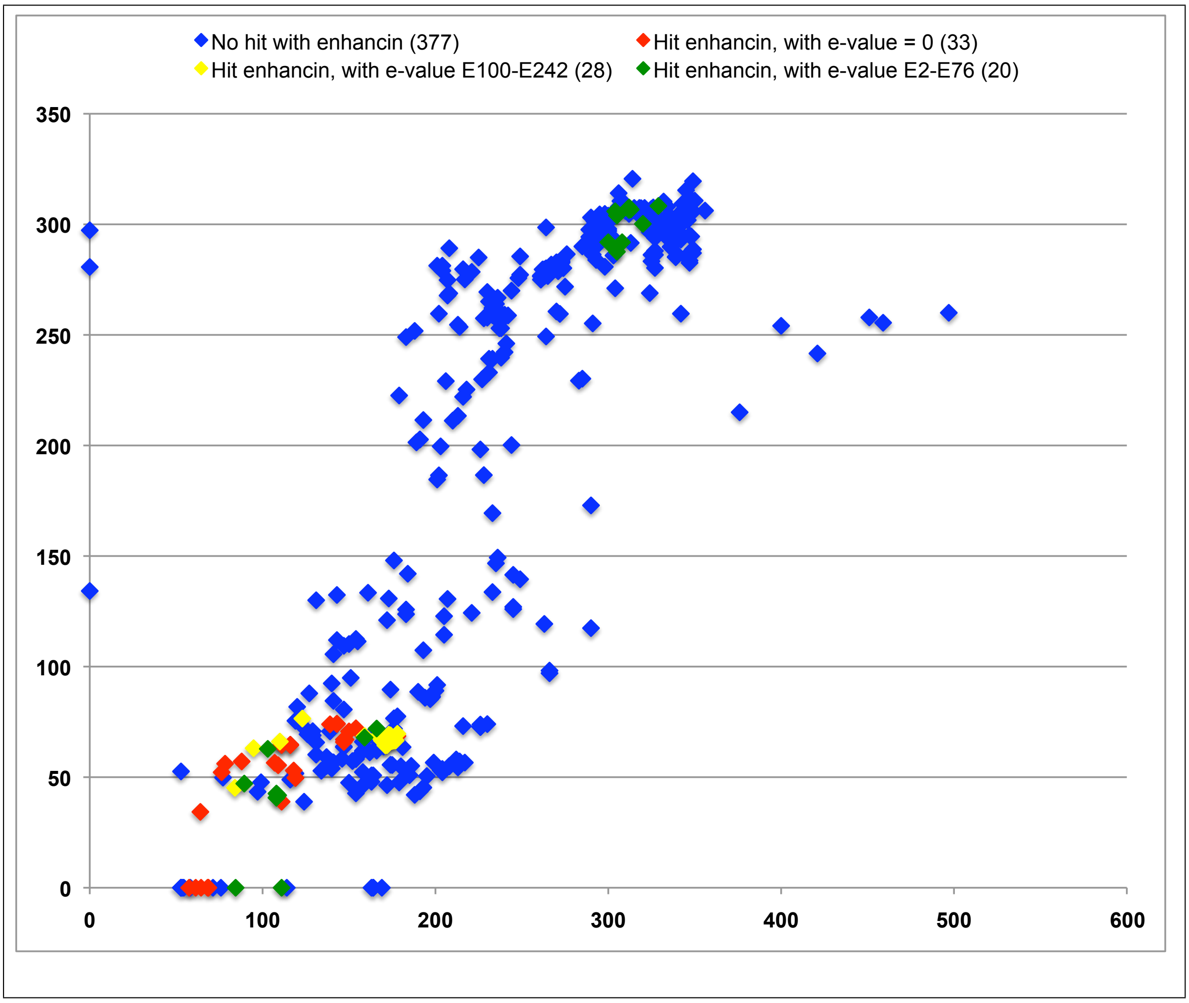


Figure 7.4: The bit scores for RefSeq proteins hit by a PSI-Blast search (e-value *le* **1**$\times10^{-5}$**) were plotted against the corresponding bit scores of the HMMER search with the M60-like profile (e-value** *le* **1**$\times10^{-5}$). HMMER and PSI-BLASTP were used to identify proteins containing an M60-like domain. The PSI-BLAST search was performed using the *T. vaginalis G3* (XP_001313628.1) sequence as a query. The scores from the HMMER search (Y axis) were plotted against hit results from the PSI-BLAST search (X axis). Sequences that are hit with M60-enhancin identified using HMMPfam are coloured based on the the e-values of the hit results. The entries without an M60-enhancin domain (i.e., no HMMPfam hits) are coloured blue. The entries with an M60-enhancin domain are coloured according to their hit e-value ranges. Numbers in brackets in the graph legend represent the total number of entries in each range. The majority of the proteins appear to have a hit to an M60-like domain and no hit to an M60-enhancin domain (blue diamonds). Some proteins are predicted to have an M60-like domain as well as a strong e-value indicating the presence of an M60-enhancin domain (red and yellow diamonds). However, these proteins have low HMMER and PSI-BLAST hits scores, suggesting a distant relation between the two domains.

domain, and animal host-associated microorganisms as well as mucosa-associated microorganisms was observed (Table 7.1). Approximately 10 bacterial taxa with M60-like positive proteins are known as free-living microorganisms or plant pathogens with no evidence for being associated with animal hosts such as *Pseudomonas syringae*, *Uncinocarpus reesii* (Appendix J). A total of 14 taxa encoding M60-like domains are animals (Appendix J) that possess mucosal surfaces such as human, cow and fish.

7.3.4 Variation of pathogenicity of Escherichia M60-like proteins family

Among 322 taxa encoding the M60-like protein domain, 32 sequences are originated from Gram-negative, non-spore forming bacteria *Escherichia* species. Three sequences are from *Escherichia sp.* (strain 1_1_43, 3_2_53FAA and 4_1_40B), which are commonly found in human intestinal tract. Twenty-seven sequences were derived from *Escherichia coli*, one from *E. fergusonii* ATCC 35469, and one from *E. albertii*. *E. albertii* and some particular strains of *E. coli* can cause infections, particularly, on mucosa surfaces of vertebrates. *E. coli* strains and *E. fergusonii* ATCC 35469 are known as members of normal gastrointestinal flora in mammals. Some strains of these species are known to be pathogenic in various mucosal niches, including the gastrointestinal, urogenital and respiratory tracts of both mammals (especially human) and avians [Kaper, 2005][Farmer *et al.*, 1985][Rasko *et al.*, 2008]. M60-like protein families are found in some *E. coli* that cause intestinal disease including enteroaggregative *E. coli* (EAEC) strain 101-1; Enteropathogenic *E. coli* (EPEC) strain E22, E110019 (atypical EPEC); Enterotoxigenic *E. coli* (ETEC) strain E24377A, B7A; Enteroinvasive *E. coli* (EIEC) strain 53638. One of major *E. coli* strains causing extraintestinal infections and which also encodes the M60-like domain is uropathogenic *E. coli* (UPEC), including strains 536, F11, and UTI89. Moreover, the M60-like protein family was also detected in *E. coli* APECO1, described as an avian pathogenic *E. coli* (APEC). The other three *E. coli* strains encoding M60-like domains are known as normal gastrointestinal microflora including strain HS, K-12 substr. MG1655, and W3110 [Blattner, 1997][Kaper, 2005][Rasko *et al.*, 2008].

7.3.5 M60-like containing protein sequences possessing carbohydrate binding modules

Several proteins containing M60-like domains possess other well-characterised protein domains and features (Table 7.2), including several features associated with cell surface or secreted proteins such as signal peptides (SP); transmembrane domains (TMD); and prokaryotic lipoprotein domains.

Table 7.2: Pfam entries annotated on proteins containing M60-like domains. Total number of each entry found is shown.

Domain description	Pfam accession	count
Coagulation factor 5/8 type, C-terminal	PF00754	88
Peptidase M60, viral enhancin protein	PF03272	83
Glycosyl hydrolase family 98, putative carbohydrate-binding module	PF08305	46
Fibronectin, type III	PF00041	38
Carbohydrate-binding family V/XII	PF02839	25
Uncharacterised sugar-binding	PF07554	19
Pyrrolo-quinoline quinone repeat	PF01011	13
Bacterial Ig-like	PF07523	11
Leucine-rich repeat	PF00560	5
Ricin B lectin	PF00652	4
S-layer homology region	PF00395	3
Surface protein from Gram-positive cocci, anchor region	PF00746	2
tRNA pseudouridine synthase	PF01416	2
Bacterial Ig-like, group 2	PF02368	2
Leucine-rich repeat, adjacent	PF08191	2
Metallophosphoesterase	PF00149	1
Dockerin type 1	PF00404	1
DNA gyrase/topoisomerase IV, subunit A, C-terminal beta-pinwheel	PF03989	1
Collagen-binding surface protein Cna-like, B region	PF05738	1

Several protein domains that function in cell adhesion or carbohydrate binding were detected in 160 proteins containing M60-like domains (see Table 7.3 and 7.4). These well-characterised domains included a galactose-binding like domain (GBD)(IPR008979:SSF49785), a coagulation factor 5/8 type, C-terminal (IPR000421:PS50022, PF00754), and the carbohydrate-binding family V/XII (IPR003610:SM00495, SSF51055, PF02839). The latter two are also annotated as CBM members of the CBM32 [Abbott *et al.*, 2008] and CBM5_12, respectively in the CAZy database[4][Park *et al.*, 2010].

Of the 98 CBM-containing M60-like proteins, 80 are from microbes that are known to colonise mammalian mucosal surfaces, including the gastrointestinal (GI) or urogenital (UG) tract. Some are well known members of the human GI tract microbiota [Gordon *et al.*, 2005][Hattori and Taylor, 2009] including *Bacteroides caccae*, *B. fragilis* and *B. thetaiotaomicron*. Several others are thought to be mainly free-living (can be isolated from the environment) but can be pathogenic when in contact with mammalian mucosal surfaces or the digestive tracts of insects. These organisms include *Bacillus cereus* [Slamti and Lereclus, 2002][Arnesen *et al.*, 2008], *Yersinia enterocolitica* subsp. enterocolitica 8081 [Thomson *et al.*, 2006], and *C. perfringens* [Brynestad and Granum, 2002].

The domain CBM5_12 or Chitin-binding domain type 3 (Pfam:PF02839, SMART:SM00495) was detected on the M60-like proteins from insect pathogens (Table 7.3) such as *Paenibacillus larvae* subsp. larvae BRL-230010 and *Bacillus thuringiensis* serovar israelensis ATCC 35646. Likewise, the co-occurrence of M60-like and CBM32 domains was identified on proteins from mucosa-associated

[4] www.cazy.org/CAZY/, accessed 10th December 2010

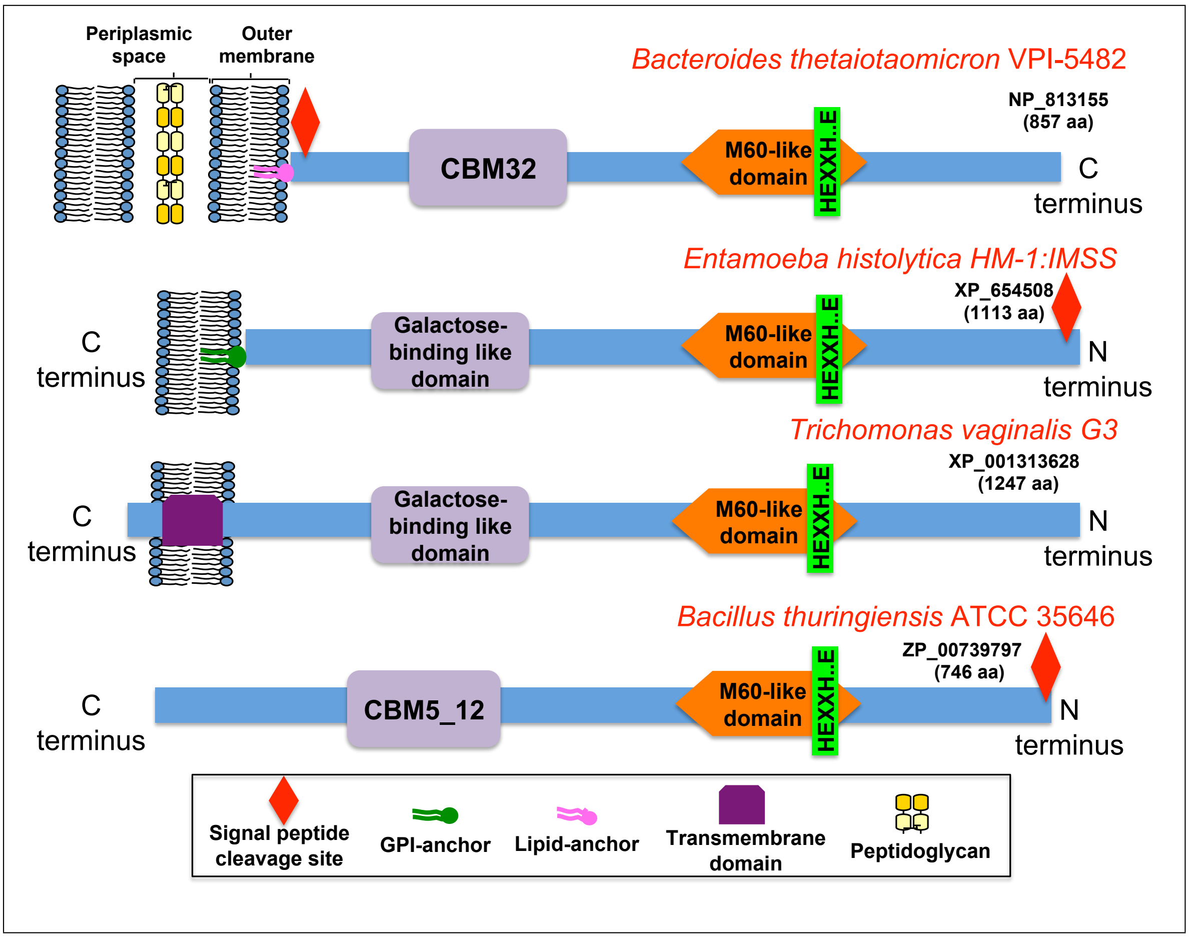


Figure 7.5: A schematic representation of the domain organisation of proteins containing both an M60-like domain and a carbohydrate-binding module (CBM) from *B. thetaiotaomicron* VPI-5482, *E. histolytica* HM-1:IMSS, *T. vaginalis* G3 and *B. thuringiensis* ATCC 35646. These microorganisms are prokaryotic and eukaryotic vertebrate mucosal microbes and insect gut pathogens, respectively. The three protein sequences have features indicating that the M60-like and CBM domains are potentially exposed to the extracellular space due to the presence of either a signal peptide or transmembrane domain. The CBM32 is commonly present in M60-like proteins derived from bacterial microbes associated with vertebrate mucosal environments, while the CBM5_12 is often present in the proteins from insect gut pathogens.

Table 7.3: List of organism species possessing M60-like proteins that have carbohydrate-binding domains from the CBM5_12 family. The number of strains and protein sequences that possess M60-like domains are shown. Several of these species are known to interact with insects (see Appendix J).

organisms	Total strain	Total sequence
Bacillus cereus[1]	9	9
Bacillus mycoides DSM 2048	1	1
Bacillus thuringiensis[2]	6	8
Bacillus weihenstephanensis KBAB4	1	1
Clostridium botulinum[3]	5	5
Paenibacillus larvae subsp. larvae BRL-230010	1	1
Yersinia aldovae ATCC 35236	1	1
Yersinia enterocolitica subsp. enterocolitica 8081	1	1
Yersinia mollaretii ATCC 43969	1	1
Yersinia ruckeri ATCC 29473	1	1

[1] 172560W,AH1134,AH603,AH621,ATCC 10876,BDRD-ST196,F65185,Rock3-28,Rock4-2

[2] Bt407,serovar berliner ATCC 10792,serovar israelensis ATCC 35646,serovar kurstaki str. T03a001,serovar monterrey BGSC 4AJ1,serovar thuringiensis str. T01001

[3] A2 str. Kyoto,B1 str. Okra,Ba4 str. 657,Bf,F str. Langeland

microorganisms (Table 7.4).

In addition, several M60-like proteins are annotated with a galactose-binding domain (GBD) which is classified as a superfamily containing the CBM32 domain. These proteins are from parasitic protozoa, *Entamoeba histolytica* HM-1:IMSS and *T. vaginalis* G3, and some are from non-pathogenic commensal microbiota of the human GI tract including *Akkermansia muciniphila*, *B. caccae* and *E. dispar*.

The CBMs are typically located on enzymes processing polysaccharides [Shoseyov *et al.*, 2006] [Boraston *et al.*, 2004]. HMMER was used to search the MEROPS database for any other known proteases also possessing either a CBM32 or a CBM5_12 domain. As a result, 182 proteins from 22 peptidase families were found to contain CBM32 domains (Table 7.5), and 33 proteins from 7 peptidase families were identified that possess the CBM5_12 domains (Table 7.6).

7.3.6 Microbial proteins with M60-like domains possess features of extracellular proteins

The majority of M60-like containing protein sequences from known mucosal-associated microorganisms were predicted to have either an N-terminal SP, TMD or lipoprotein. The presence of a SP or lipoprotein suggests extracytoplasmic localisation, while a TMD enables the anchoring of a protein to cell membrane lipid bilayers. Topological inference for the putative transmembrane proteins encoding M60-like domains, indicates that the domains are exposed to the extracellular milieu.

Table 7.4: List of organism species possessing M60-like proteins that have carbohydrate-binding domains from CBM32 family. Number of strains and protein sequences that possess M60-like domains are shown. Several of these species are known as human gut normal flora. Some are regarded as pathogenic to vertebrate hosts via the host mucosa surfaces (see Appendix J).

organisms	Total strain	Total sequence
Bacillus cereus[1]	10	10
Bacillus thuringiensis[2]	4	6
Bacteroides caccae ATCC 43185	1	9
Bacteroides finegoldii DSM 17565	1	1
Bacteroides fragilis[3]	3	3
Bacteroides plebeius DSM 17135	1	1
Bacteroides sp.[4]	3	5
Bacteroides thetaiotaomicron VPI-5482	1	4
Clostridium bartlettii DSM 16795	1	1
Clostridium botulinum[5]	2	2
Clostridium difficile QCD-32g58	1	1
Clostridium hathewayi DSM 13479	1	1
Clostridium hiranonis DSM 13275	1	1
Clostridium perfringens[6]	9	15
Clostridium sp. 7_2_43FAA	1	1
Eggerthella lenta DSM 2243	1	1
Eubacterium dolichum DSM 3991	1	1
Sphingobacterium spiritivorum[7]	2	4
Trichomonas vaginalis G3	1	2

[1] AH1134,AH676,ATCC10876,B4264,m1550,MM3,Rock1-15,Rock1-3,Rock3-28,Rock 3-29

[2] Bt407,IBL200,serovar beliner ATCC10,serovar thuringiensis str. T01001

[3] 3_1_12,NCTC9343,YCH46

[4] 1_1_6,2_1_16,3_2_5

[5] E1 str. 'BoNT E Beluga',E3 str. Alaska E43

[6] ATCC 13124,B str. ATCC 3626,C str. JGS1495,CPE str. F4969,D str. JGS1721,E str. JGS1987,NCTC 8239,SM101,str. 13

[7] ATCC 33300,ATCC 33861

Table 7.5: MEROPS entries annotated with CBM32 domains. The total number of MEROPS entries found for each protease family are shown.

Protease family	Bacteria	Eukaryote	Total
C01A		2	2
C02A	1		1
I01		9	9
I08		15	15
I43		3	3
I63		28	28
M04	1		1
M06	4		4
M12A		1	1
M12B		6	6
M14B		70	70
M14X		2	2
M20A	6		6
M23B	3	3	6
M36	3		3
M60	1		1
M64	1		1
S01A		4	4
S08A	9	1	10
S45	4		4
S63		4	4
T06		1	1
Total	33	149	182

Table 7.6: MEROPS entries annotated with CBM5_12 domains. The total number of MEROPS entries found for each protease family are shown.

Protease family	Archaea	Actinobacteria	Firmicutes	Proteobacteria	Total
M04		3			3
M06			2		2
M28A		2			2
M60			1		1
M64				1	1
M66				1	1
S01A		9		5	14
S08A	2		1	5	8
S53		1			1
Total	2	15	4	12	33

Therefore, M60-like domains are likely to be presented on the cell surface or are secreted and therefore interact with extracellular substrates, either from the host or other members of the microbiota. Prokaryotic membrane lipoprotein lipid attachment domains (PROSITE:PS51257) were detected in some M60-like proteins from free-living microorganisms, often found to be toxic to the host through the mucosal surfaces e.g. *Vibrio parahaemolyticus*, *V. cholerae*. In contrast, extracellular-associated sequence features were not detected on M60-like proteins from any eukaryotic hosts, known plant pathogens (*Pseudomonas syringae*), and some animal pathogens such as *Yersinia pseudotuberculosis*.

7.4 Discussion

7.4.1 M60-like as a potential zinc metalloprotease and enhancin-related protein family

A range of evidence, when considered together, strongly supports the hypothesis that the M60-like domain is a new metalloprotease. Firstly, the presence of the extended consensus HEXXHX(8,28)E motif, suggests that the M60-like domain could be considered as a gluzincin metalloprotease. Bacterial and mammalian gluzincins include thermolysins, endopeptidase-24.11, angiotensin converting enzymes and aminopeptidases, with the length of an inserted region between the second H and E ranging from 24-64 amino acids [Hooper, 1994]. However, none of the consensus sequences of these gluzincins peptidases correspond to the consensus region found among the proteins possessing M60-like domains. Secondly, using profile-profile comparisons, the M60-enhancin protease family was detected as a remote homologue to the M60-like profile. Although it is difficult to predict functions from distant homologous protein sequences, the structural similarity of the two protein families can be inferred [Söding *et al.*, 2005]. However, to date, no three-dimensional structure of any member of the M60-enhancin family has been produced. Enhancins are enzymes that degrade mucin-like substrates and were originally discovered in *Trichoplusia ni* baculovirus and granulovirus proteins [Wang and Granados, 1997]. Enhancins were shown to promote viral infection in the lepidopterous insect. The viral enzymes have been shown to have a degrading activity both *in vivo* and *in vitro* and were classified as metalloproteases (Family M60, clan MA; subclan MA(E)) possessing the classical HEXXH motif [Wang and Granados, 1997] [Lepore *et al.*, 1996]. Taken together, these data strongly support the hypothesis that the M60-like domain represents a new gluzincin zinc-metalloprotease. In addition, the M60-like candidate metalloproteinase can be classified into clan MA, subclan MA(E) according to the MEROPS peptidase classification schema by both molecular structure and homology.

7.4.2 Carbohydrate-binding domains on proteins possessing M60-like domains

For vertebrates, mucus layers produced by epithelial cells are a physical surface barrier facing the external environment of several organs such as the GI tract, respiratory tract and UG tract [Nagler-Anderson, 2001] [Vélez *et al.*, 2007] (see Figure 2.2). Similarly, the invertebrate digestive tract also possesses a protective mucin-like layer and in the insect gut this is called a peritrophic membrane. Unlike vertebrate mucins, a major component of invertebrate peritrophic membranes is a chitin rich matrix [Wang and Granados, 1997]. Both vertebrate and invertebrate barriers play important roles in protecting the digestive tract from microbial infections, as well as promoting digestion processes [Turnbaugh *et al.*, 2006][Flint *et al.*, 2008]. Therefore, in order for a microbe to colonise or penetrate these protective barriers, physical interactions and enzymes capable of processing these components are required.

Several proteins possessing M60-like domains encoded by insect pathogens contained a carbohydrate-binding module family V/XII (CBM5_12;CMB5 and CBM12). The CBM5_12 domains are defined as chitin-binding modules and are found mainly as components of bacterial chitinases and other different carbohydrate degrading enzymes [Brun *et al.*, 1997][Ikegami *et al.*, 2000]. Several insect-infecting pathogens encode chitinases to penetrate through the chitin barriers [Wang and Granados, 1997] [Abbott *et al.*, 2008] [Sampson and Gooday, 1998]. The presence of a C-terminal CBM5_12 on the M60-like proteins of the Gram-positive, spore-forming insect pathogens *P. larvae* and *B. thuringiensis* suggests that these proteins are able to bind to the chitin-rich peritrophic membrane and to degrade protein components through the M60-like potential peptidase. The M60-like-CBM5_12 proteins are also predicted to possess a SP, suggesting these proteins might be facing the extracellular space either as a cell surface or secreted protein.

P. larvae is a causative agent for American foulbrood (AFB) disease of honeybee larvae. Young bee larvae are susceptible to the infection by ingesting spores from virulent strains of *P. larvae*. The spores geminate in the gut of the bee larvae and cause disease in the larvae host [Qin *et al.*, 2006]. Metalloproteases were reported to be involved in the pathogenicity of AFB [Antúnez *et al.*, 2009]. The predicted extracellular M60-like-CBM5_12 protein from the bee pathogen has chitin adhesion abilities. The M60-like domain on this protein also contains a HEXXH. . . .E gluzincin metallopeptidase motif, suggesting that the protein is a potential virulence factor involved in bacteria-insect interactions.

While, CBM5_12 was a feature found on the insect's proteins carrying M60-like domains, carbohydrate-binding module family 32 (CBM32) sequences were detected on many of vertebrate pathogens encoding M60-like proteins. CBM32s are found in wide range of microorganisms, particularly, plant and animal pathogens. Ligand targets of the CBM32 range from plant polysaccharides to eukaryotic complex glycans [Abbott *et al.*, 2008].

Surprisingly, this study reveals CBM-like sequences linked to a predicted zinc-metallopeptidase rather than carbohydrate-active enzymes [Boraston *et al.*, 2004]. These findings reveal a new functional context of CBMs. Their role is likely to enable the attachment of peptidases to glycoproteins, such as host mucosal surface barriers, thus contributing to the ability of microbes to attach to, and degrade, host mucins.

In addition, some experimental and microarray data has been reported on two M60-like proteins from *V. cholera* and *B. thetaiotaomicron* VPI-5482, respectively. Hughes *et al.* [Hughes *et al.*, 1994] have shown that a mutation of the *V. cholera acfD* gene, that encodes an accessory colonization factor AcfD precursor (M60-like protein), dramatically decreases the microorganism's motility. Based on microarray data, the M60-like protein (BT_4244) from the gut-derived *B. thetaiotaomicron* VPI-5482 is upregulated when the bacterial cells are exposed to mucin [Martens *et al.*, 2008]. The *B. thetaiotaomicron* protein is encoded by a gene that is known to be part of the bacterial starch utilisation system (SUS) [Martens *et al.*, 2008].

7.5 Conclusions

A novel protein domain, named M60-like, was identified and defined. The domain represents a potentially novel family of extracellular metalloproteases that is hypothesised to play an important role in animal host-microbe interactions. The M60-like domain is shared across a broad range of prokaryotic and eukaryotic mucosa-thriving symbiotic and pathogenic microorganisms, as well as mucosa-possessing eukaryotic hosts. Mucosal niches for microbes encoding M60-like domains include the human urogenital tract, vertebrate gastrointestinal tract and respiratory tract and the insect gut. *In silico* characterisation of proteins possessing M60-like domains derived from mucosal microorganisms revealed a possible novel proteolytic activity ascribed to the domain. The conserved HEXXH(8,28)E motif and the relationship of the M60-like HMM profile to the enhancin domain indicate that the proteins are gluzincin zinc metalloproteases. Moreover, several lines of evidence suggested that extracellular localisation of the M60-like proteins. A subset of the microbial M60-like domain-containing proteins can further be characterised by the presence of the CBM32 or CBM5_12. The proteins containing M60-like-CBM32 domains were mainly encoded by the genomes of microbes dwelling on vertebrate mucosal surfaces, including important commensals and pathogens. In contrast, M60-like-CBM5_12-containing proteins were detected in insect-infecting bacteria. A novel functional context for CBMs was also identified, which are typically connected with carbohydrate-processing enzymes but not proteases. The CBM domains linked with proteases are likely to enable various proteases to bind to specific glycoproteins from host mucosal surfaces (e.g. mucus, glycocalyx), further highlighting the importance of CBMs and proteases in host-microbe interactions. In conclusion, the M60-like domain may be involved in a specific host-microbe interaction processes. Mucosal microbial surface proteins play multiple essential roles in initiating and sustaining the colonisation of the heavily defended mucosa. The M60-like domain may play roles in adhesion, degradation of the ECM or peritrophic membranes, invasion, or killing of host immune cells. The identified structure features of the M60-like protein highlight this protein as a candidate for future laboratory studies addressing their function importance in the context of mucosa-microbe interactions and colonisation. The results described in this chapter are a good example of what can be achieved through detailed bioinformatics analyses for the purpose of hypothesis generation regarding the functionality of uncharacterised proteins. The analysis performed in this chapter could be applied to a range of protein domains of unknown function or uncharacterised conserved protein regions that were identified in Chapter 6.

Chapter 8

Conclusions, Discussion and Future Work

8.1 Overview of different aspects of the project

8.1.1 A high-throughput sequence analysis workflow

This study has demonstrated the application of Grid and Cloud technologies to bioinformatics workflows for the analysis of 867 microbial proteomes. The high-throughput workflows performed a large-scale analysis of 3,021,490 protein sequences in order to predict extracytoplasmic proteins, detect sequence signatures and search for sequence similarity. The extracytoplasmic protein prediction pipeline was designed to process proteomes from different cell surface structures (e.g. archaea, Gram-positive and Gram-negative bacteria, as well as microbial eukaryotes) by automatically applying a selected set of prediction tools and strategies.

The workflow was developed using Microbase and is fully automated with minimal human effort required for the installation of bioinformatics software, computational task distribution and execution, and result compilation and storage. Moreover, the workflow can be reused and could potentially be modified or extended to facilitate other requirements of further analyses. In deed, a number of components of the sequence analysis workflow developed during this study are currently actively used in another research project (AptaMEMS-ID) [McNeil *et al.*, 2010], for the identification of unique surface proteins of infectious microorganisms such as *Staphylococcus spp*.

8.1.2 Microbe-habitat annotation

Information about organism habitats or isolation sources was required in order to identify functional features specific to mucosa-thriving microbes. There is currently no comprehensive habitat information available in public databases. The richest source of accessible organism-habitat information is published literature. However, free-text publications are not immediately computationally-accessible for large-scale comparative genomics studies and the collection of relevant scientific literature is too large for manual annotations to be feasible. Therefore, exploitation of this information requires the development of an automated annotator tool, based on a text-mining approach. In order to standardise vocabularies referring to the habitats of microorganisms, a prototype habitat ontology designed to describe environmental and host-associated habitats was constructed. This ontology focused on providing definitions of animal anatomical niches, particularly mucosa-lined cavities, and other environmental habitats.

8.1.3 Comparative genomics and the identification of the genotypic features overrepresented to mucosal microorganisms

The comparative genomics study performed in this study revealed a contrast in the distribution of known conserved protein domains between known mucosa-thriving and microorganisms from other habitats. The genotypic features associated with a mucosal environment were identified by testing the statistical significance of either the co-occurrence or the abundance of those elements among microorganisms annotated to thrive on mucosal surfaces. Some of the identified protein domains correspond to known to be involved in promoting the survival, or aiding pathogenicity of microbes in the highly-defended mucosa. However, a number of protein domains with unknown function were identified that could potentially play various important roles in the complex interaction of microbes and the host mucosal environment.

Several groups of homologous extracytoplasmic proteins from known mucosa-thriving microbes were identified within and between prominent gut commensals (Bacteriodes, Firmicutes and Proteobacteria) as well as bacteria and microbial eukaryotes from other mucosal surfaces. Many of these extracytoplasmic protein families do not possess any previously identified conserved region, and in most cases their functions are unknown. The patchy taxonomic distribution of both the identified candidate mucosa-associated protein domains and protein families suggests that lateral gene transfer of these genetic elements played an important role in the evolution of mucosa-associated microbes.

Based on the identified mucosa-associated extracytoplasmic protein domains with known functions, these elements appear to be involved in several processes. These processes include carbohydrate or amino acid transports, metabolic processes, attachment to host tissues or other substrates in the environment, signal transduction and cell communication, and resistance to host defence mechanisms. For pathogenic strains, these functions also include various invasive and virulence factors.

8.1.4 A novel host-associated catalytic protein domains

Finally, a novel protein domain, named M60-like, was identified and deposited in Pfam database (accession PF13402). The M60-like domain was shown to be encoded by animal hosts and host-associated microorganisms. These microorganisms include insect-related and mucosa-associated commensals and pathogens. Detailed bioinformatics analyses have suggested that proteins possessing M60-like domain are new candidate extracellular proteases that could assist microbial survival and colonisation on mucosa surfaces. A potential catalytic function of the conserved gluzincins metalloprotease motif was found as part of the M60-like domain. These analyses also identified the co-occurrence of the M60-like domain and the carbohydrate-binding modules (CBMs) on the same protein sequences, revealing a new functional context for the CBMs, which are typically connected with carbohydrate processing enzymes, but not proteases. This finding further emphasise the importance of extracellular proteases and CBMs in host-microbe interactions.

8.2 Discussion

8.2.1 Advantages and challenges in using the high-throughput analysis workflow

The sequence analysis workflow developed during this project has served the needs of a large-scale comparative genomics study. In this project the workflow was used to identify microbial extracytoplasmic elements, from microbiota that interact with host mucosa surfaces. Such an extensive genotype-habitat correlation analysis has not previously been performed for mucosal microbes, due to existing analysis capacity limitations and the lack of good quality (and sometimes non-existent) habitat annotations of taxa. Therefore, this study provides new insights into the biological significance of the genotypic features important for microorganisms to successfully thrive in host mucosal environments.

The workflow is highly automated, a feature that facilitates the analysis of large data sets, but also brings new challenges. Errors in the system, such as when incorrect or corrupt input data automatically retrieved from public databases is given to the system, can be difficult to detect. Automated analysis can also produce some errors that may not be detected easily given the scale of data. To address these issues, automated systematic cross-checking methods were implemented on various parts of the project's workflow output on a regular basis.

Using Microbase reduces the time required for processing large amounts of data. The use of a processing pipeline reduces human intervention, allowing comparative genomics studies to be performed in an automated fashion. Analysis of the protein sequence data across all three domains of life was completed within a reasonable time frame (three months), providing up-to-date analysis results based on data available at that time. In total, five years of compute time was used. New and updated genome sequence data can be incrementally added into the workflow, facilitating further dynamic comparative studies when new sequence data becomes available.

The ability to chain together any project-specific functionality to form an analysis pipeline that can be processed in a distributed computing system provides a framework for an automated large-scale genomics analysis in the post-genomics era. The analysis pipeline developed for this project may be used as a base from which other projects can extend and enhance its functionality in the future.

8.2.2 Challenges in identifying microbial eukaryotic extracytoplasmic proteins

The high-throughput extracytoplasmic protein prediction pipeline developed during this study represents a novel set of workflows that integrates existing targeting-signal prediction tools for the analysis of protein sequences from all three domains of microbial life. The pipeline efficiently identified prokaryotic extracytoplasmic proteins from primary sequence data. However, the performance of the system in the prediction of microbial eukaryotic proteins has not yet been investigated due to the difficulty in finding accessible experimentally-verified subcellular localisation data for microbial eukaryotic protein data sets, as well as the complexity due to different cellular organisations. Furthermore, the computational identification of extracytoplasmic proteins of microbial eukaryotes is challenging, as these organisms have complex endomembrane systems with many distinct organelles. Therefore, targeting signals or anchoring features could potentially be found on proteins derived from and specific to these organelles.

8.2.3 Statistical analysis for the genotype-phenotype association

The bioinformatics approach used in this study allows an exploration of microbial components across different environments through the use of existing genome data and the available information about their natural habitats or sources from which they were isolated. At the time of conducting this study, availability of isolation source information of complete genome sequences and proteomes was limited. The comparative genomics and multivariate analysis used in this study have proved to be valuable approaches. The use of a hypergeometric distribution as a significance test has resulted in a meaningful list of conserved elements known to be important for the survival of microbes in a particular ecological niche. Large numbers of candidate mucosa-associated proteins and protein domains were identified. Several of these protein-coding genes' functions are unknown or uncharacterised. This work therefore provides a list of candidate genes to prime further investigations, both computational and biological. The investigation provides an insight into the understanding of host-microbe interactions from the microbial genomics perspective. These new insights might provide opportunities to develop new probiotic and prebiotic substances as well as new therapeutic agents [Jia *et al.*, 2008][Sekirov *et al.*, 2010].

The quality of the results in terms of sensitivity and specificity is anticipated to increase as more precise and complete information about habitat or isolation source of microorganisms becomes available. A more detailed repository of microorganism habitat data can be acquired by extracting the relevant information from the published literature. The initial training of a machine learning approach for the development of automated text-mining tools capable of capturing microbe-habitat pairs was initiated during the course of this study. This work fulfils the need for developing the automatic capture of metadata referring habitats of microorganisms, to serve the growing number of genomics and metagenomic projects [Hirschman *et al.*, 2008]. Another aspect to be considered in parallel with the metadata mining is the standardisation of vocabularies as control terms representing different (ecological) properties of microorganisms. The prototype microorganism habitat ontology developed in this study, together with the automatic capture of microbe-habitat information aids meaningful comparisons for studying the associations between habitat and genotypic features. Additionally, the ontology and text-mined metadata together should enable a systematic analysis in a comparative genomics study where large amounts of available genome data are included and habitat information is the key focus of the research questions. A clear example can be seen in our study where a more thorough association analysis of genetic materials and ecological properties of microorganisms can be performed. In particular, the approach could be applied to more specific mucosa-lined niches, such as a comparison between 'colon' and 'urogenital tract', as well as among other potentially non-mucosal ecological niches. The study could be conducted effectively once high quality of habitat information are obtained.

8.2.4 Choices of statistical methods

Several large-scale research programs have been initiated to investigate the human microbiome with a current emphasis on the gut (e.g. [Martens *et al.*, 2009][Ellrott *et al.*, 2010][Hattori and Taylor, 2009]), resulting from the availability of initial sequence data from the Human Microbiome Project (HMP) [Turnbaugh *et al.*, 2007][Consortium *et al.*, 2010] and Metagenomics of the Human Intestinal Tract (MetaHIT) consortium [Qin *et al.*, 2010]. However, the study described in this thesis represents an expanded analysis of all vertebrate mucosa microbial communities, mostly the human microbiome and pathogens, rather than being restricted to specific areas of the body. The statistical approach as well as the analysis workflow can potentially be used to integrate existing complete genome data with new sequence data from the metagenomics projects once it becomes available. Sequence data from a metagenomics project will provide accurate information about the isolation source of the sequences which should allow the comparative study of more specific body sites within or between individuals.

During the course of this project, several statistical approaches were investigated for the purpose of correlating functional genetic elements to the ability of microbe to survive in particular environments. The techniques investigated included both bivariate and multivariate data analysis techniques. Each technique is discussed in terms of its suitability to the project's data, and its advantages and disadvantages from a practical perspective. The statistical techniques considered for this project were Pearson's correlation analysis, Hypergeometric distribution test, Propensity scoring, Mutual Information (MI), Step-wise discriminant analysis (SWDA) and Principal component analysis (PCA). Each of these techniques were applied to a subset of the data set to evaluate and learn the process of applying the techniques and the biological meaning of the results obtained from different techniques.

Pearson's correlation coefficient was used to evaluate the correlations between the identified protein domains and the ability of microorganisms to thrive on a mucosa-lined habitat. The dataset for this use case was a set of 11 lactic acid bacterial genomes: 5 isolated from human gut environment; 3 from diary products and 3 found to present in multiple niches. This initial data set is the same data set used by O'Sullivan and colleagues [O'Sullivan *et al.*, 2009]. InterproScan results of all protein sequences from the 11 bacterial genomes were obtained from our analysis workflow. Pearson's correlation scores were calculated to find the correlation between the mucosa-related lifestyle of the human-gut microbes and well-characterised protein domains from various domain databases. The results of protein domains showing positive correlations with the gut-isolated lactobacilli corresponds to the previous study revealing proteins involved in the phosphotransferase system (PTS) and the glycosidase system are found to present in the gut microbes [Lozupone *et al.*, 2008]; the Pfam domain

PF00367 and PF03611 with correlation scores = 0.54 and 0.60 respectively. The domains PF02903, PF00128, PF02449 with scores 0.62, 0.58 0.56 respectively–involved in glycoside hydrolase activity which has also been previously reported to be present in gut microbiome [Jim, 2003]–were also identified by the Pearson's correlation analysis approach. The PF02275 domain was also shown to correlate with a gut-mucosal thriving ability with a score = 0.59. The PF02275 domain, having a choloylglycine hydrolase activity, is found to be present on the protein products of the proposed bile-salt hydrolase gut-specific genes [Liu *et al.*, 2006]. From this result, applying Pearson's correlation analysis to correlate functional protein domains and the mucosa-related life style of microorganisms appears to be a promising approach for this project.

However, measuring the strength of the correlation between the M60-like domain and a microorganisms' ability to thrive on mucosa-lined niches by calculating Pearson's correlation yielded a weak positive correlation (score = 0.21). This weak correlation is due to the large number of taxa annotated as mucosal that are not positive for the M60-like profile. This result suggests that the use of Pearson's correlation analysis is not always useful for this type of biological pattern, where the conserved functional region can be shared among a very restricted set of the interested taxa.

For the distribution pattern of the M60-like domain, there is a high proportion (45/62, approximately 73%) of mucosa-thriving microorganisms among all microorganisms with M60-like-containing protein sequences. For this reason, the propensity score (Φ) [Jim, 2003] (see Section 2.9.1) was calculated as an alternative measurement to quantify the association between the microorganisms' mucosa-thriving ability and the M60-like domain. The propensity score for the M60-like domain is 1.7, whereas the maximum propensity is 2.3. The maximum propensity score is used as a maximal base line to determine the strength of the genotype-phenotype correlation. The propensity score of the M60-like domain and mucosa-associated lifestyle is 74% of the maximum propensity score, meaning that 74% of microbes possessing M60-like domain hit protein sequences are mucosa-thriving microorganisms. The association is statistically significant with a p-value of 3.7×10^{-9} (calculated using hypergeometric distribution function).

In brief, using the Pearson's correlation scores to correlate organisms' gene-derived protein products to phenotypes does not always indicate a strong correlation if a given genotype does not have a significant linear relationship to the phenotype and is not distributed across all taxa with the given phenotype. The propensity score allows the identification of the association between the phenotype and a genotype that are conserved within a subset of taxa displaying a given phenotype. However, using the hypergeometric distribution also provides a significant confidence level, so the overrepresence of the M60-like domain in mucosal microorganisms is unlikely to happen by chance.

Mutual Information (MI) scoring (see Section 2.9.2) is another method that was investigated during the course of this study. The technique was used by Slonim *et al.* [Slonim *et al.*, 2006] to measure correlations between genes and observed phenotypes of organisms. This technique was shown to provide biologically meaningful results in the study by Slonim *et al.*. In our case, the equation provides a measure of whether each protein domain correlates with microorganisms known to be able to thrive in mucosal environments. The MI score for each domain was calculated with the following input values:

- the number of taxa annotated as mucosa-dwelling, and have that domain;
- the number of taxa annotated as mucosa-dwelling but do not have that domain;
- the number of taxa annotated as non-mucosa-dwelling and have that domain;
- the number of taxa annotated as non-mucosa-dwelling and do not have that domain.

The equation (see Section 2.9.2) was used to calculate scores for each domain. However, by the nature of its formula, the technique is not applicable where one or more of the input values is zero. The log operation will result in infinity if one or more of the inputs is zero, meaning that the resulting score is of no use. When all of the inputs are non-zero, applying MI to the data set provides similar results to using the hypergeometric test. However, using the MI on a subset of data demonstrated that in several cases, an input values are zero. This suggests that MI might not be suitable for estimating correlation of protein domains and organism's phenotypes.

The data set used in this project contains multiple variables. This project was concerned with finding patterns of relationships between these variables. Multivariate statistical techniques such as clustering algorithms, principle component analysis (PCA), and discriminant analysis (DA) were also investigated during the course of this study.

Clustering analysis techniques were shown to be useful methods in the analysis of the data set in this project since protein domains can be grouped together based on their pattern of distribution across organisms. Protein families can then be visualised and summarised in a straightforward manner in relation to the isolation source of organisms. Likewise, protein homologues can be clustered together based on their similarity scores. The clustering techniques have proved a valuable method for summarising large amounts of data into a succinct view making it more amendable to make an interpretation. Such technique allows a greater understanding of the complex biological relations where multiple variables are involved and biological means were searched for.

PCA is one of the methods that can be used to reduce the dimensions of a multi-dimensional data set. PCA uses correlations among the variables to develop a small set of components that empirically summarise the correlations [Tabachnick and Fidell, 2001]. Variables that correlate to each other but largely independent from other subset of variables are combined into components. PCA allows a better visualisation of the overall patterns of variation or correlation among variables by reducing a large number of observed variables to a small number of components [Tabachnick and Fidell, 2001]. PCA was applied to a subset of a small set of data generated in this project to explore the use of PCA for our study. The aim was to use PCA to reduce roughly 4,000 protein domains (variables) down to a much more manageable numbers of components that summarise a pattern of correlations among the observed variables. However, the components produced by PCA reflect the nature of the underlying processes forming the correlations among variables, which may not be the same processes that are of interest to the researcher. This project was concerned primarily with the processes underlying the ability of microbes to thrive in different habitats. Unfortunately, the author found that the first few components do not account for the most percentage of the variance in the observed variables. The results obtained do not represent a good principle component analysis, where a high percentage of the variance should be accounted for by the first few components [Tabachnick and Fidell, 2001]. The result also suggested that the data set of protein domain profiles from wide range of microorganisms contains a high level of noise. The high volume of noise in the data set is likely to be explained by the complex interactions of microorganisms' genotypes and phenotypes. For example, there are a number of combinations of protein domains that are important for the survival of microorganisms on mucosal environments and organisms may use different strategies for their surviving in the same environment. Moreover, the project involved different taxonomic groups.

Step-wise Linear discriminant analysis (SLDA) is a statistical analysis technique that can be used to indicate variables that discriminate the differences between a number of groups of dependent (categorical) variables. In our case, two-group discriminant analysis can be employed to find linear combinations of the protein domains (discriminant variables) that enable the separation of mucosa-thriving and non-mucosa categories. The SLDA technique not only determines which variables might be involved in group discrimination, but it also determines which variables account for greatest differences in the mean profiles of the two groups. The SLDA test is currently only available in the SPSS statistical package, Mac OSX and Windows platforms. This limitation made it impossible to run SLDA on the servers available to the author, since these servers use Linux. Running the set of approximately 4,000 protein domains (variables) from 400 organisms (subjects) takes over 2 GB of memory, which was beyond the capability of the available hardware. Due to this practical difficulty, the SLDA technique was not used for this study.

Another tool, the Targeted Projection Pursuit (TPP) [Faith, 2007] was also considered in this study. The TPP can be used to perform exploratory discriminant analysis where users can explore the relationships between dependent and independent variables. TPP provides an intuitive graphical interface allowing users to explore high-dimensional data by manipulating the view of data. The tool calculates linear projections of data using PCA and plots the nature structure of data in two-dimensional scatter plot. The users can dynamically explore other possible views or the structure of the data that they are interested in. The linear projection that produces the view that best matches the user targeted view can then be found. TTP appears to be a promising tool for the exploration of complex data sets in order to identify the projections that discriminate each microorganism habitat from each other where several organisms with different habitats and protein domain profiles are involved. However, the current version of this tool is not capable of handling the large amount of data such as the data set in this study.

8.2.5 Primary sequence analysis results and data integration

The GenomePool database provides a query-able data source of genomes and their corresponding annotations. A set of genes or proteins of interest can be accessed and obtained programmatically. The database allows the interconnection between the genomics information and the organisms' phenotypic description. The analysis results from each bioinformatics tool used in the project can be integrated providing a comprehensive set of information about protein-coding gene sequences of various microorganisms.

The pre-computed InterProScan results of a wide range of microorganisms where results are query-able are also of great interest to microbiologists. The domain composition of their favourite organisms can be generated in a straightforward manner. Domains of interest can also be queried for their distribution among taxa. Corresponding protein-coding gene sequences can then be identified for further investigation.

In the same way, the putative set of microbial extracytoplasmic proteome data generated by the project extracytoplasmic protein identification workflow provides a good pre-computed resource for a rapid prediction of protein localisation based on the results from several well-known and widely used targeting-signal prediction tools.

8.3 Future work

Genome annotation is a dynamic process. New genome sequences are released on a daily basis, while the annotations of previously published sequences are improved over time. It would be interesting to rerun the analysis on any updated or new genome sequence data as they are made available. The automated nature of the system developed in this work means that the analysis can be repeated in a relatively straightforward manner.

The high-throughput workflow discussed in Chapter 3 provides large-scale protein sequence analysis functionality. The statistical analysis (such as those described in Chapter 6) of the generated data from the workflow was performed with minimal automation. Therefore, adding new genome sequences to the data set is time consuming since a degree of manual statistical analysis process is required. A further enhancement to the project would be to integrate these statistical analysis processes with the workflow. By automating the entire process, the list of proteins or domains potentially involve in the microbe-mucosa interactions could be more easily updated when new genome sequence data became available. An up-to-date list of candidates takes consideration more sequence data, should provide more reliable results to guide experimental analyses in order to find novel strategies for promoting our health and preventing diseases.

The performance of the extracytoplasmic protein identification workflow may be improved by adding prediction tools specific to the unusual targeting signals for non-classical secretory pathways of some groups of microorganisms such as Archaea, Mycobacterium and Gram-negative bacteria. For example, the type III secretion system (TTSS) is one of the Gram-negative bacterial secretory systems known to mediate interaction between host cells and pathogenic bacteria [Deng *et al.*, 2004] [Zumaquero *et al.*, 2010] [Spinner *et al.*, 2008]. Tools specific for the prediction of: archaeal signal peptide [Bagos *et al.*, 2009] [Yu *et al.*, 2010], Mycobacterium tuberculosis secreted proteins [Gomez *et al.*, 2000], the type III secreted proteins [Arnold *et al.*, 2009] and beta-barrel transmembrane proteins [Bagos *et al.*, 2004b] [Natt *et al.*, 2004] can be added to the pipeline. The effect of using the combined transmembrane and signal peptide prediction tools, such as Phobius [Kall, 2004], can be explored in relation to the final list of putative extracytoplasmic proteins.

The extension of the workflow to classify extracytoplasmic proteins into their targeted locations is beyond the scope of this project. However, this extension of functionality would provide a great benefit for the large-scale prediction of microbial protein subcellular localisations. Tools such as PSORTb [Yu *et al.*, 2010], LocateP [Zhou *et al.*, 2008] and BaCelLo [Pierleoni *et al.*, 2006] can be added to form another workflow for the purpose of protein subcellular location prediction.

Another improvement that could be made is the inclusion of the latest version of InterProScan into the domain recognition workflow. The new versions contain the latest collections of protein domain databases as well as faster and more efficient algorithms. Moreover, from InterProScan version 4.5 onward, the tool has integrated the High-quality Automated and Manual Annotation of microbial Proteomes (HAMAP) database [Lima *et al.*, 2009].

The author has overseen the development of the necessary infrastructure for building and maintaining a large-scale integrated database. This database combines data from many distinct sources. For example, it is possible to query any protein-coding genes present in the GenomePool for InterProScan hits, and then cross-reference these with results from other targeting signal prediction tools, such as SignalP, LipoP and TMHMM. Throughout this project, the database has proven invaluable for such integrated queries. In the future, it should be feasible to expose this database as a public resource, allowing biologists and bioinformaticians to perform remote queries. A Web interface could be provided for manual data browsing, while a Web Service interface could be made available for programmatic access to the data.

Even though the system for mining biomedical literature developed during the course of this study did not proceed beyond a prototyping stage, the system was trained to automatically ascribe microorganisms to their habitats, where it performed well with small data set. With additional development and further fine tuning and testing with a larger data set, it should be possible to obtain highly informative habitat annotations. This is currently being done by the collaborator in Manchester university. Rule-base systems would then be applied to manipulate habitat terms into the knowledge-based project-specific habitat ontology to allow a systematic identification of microorganisms' habitats. Then, statistical analyses can be re-performed for the identification of genotypic features important to mucosa-thriving microbes. The more detailed habitat annotation would allow more fine-grained genotype-habitat analysis. For example, the contrast of genotypic features from various mucosa-lined niches can also be investigated such as gut, oral cavity, and urogenital tract. Likewise, a comparative genomics study of microbial genomes found in different host species could then be performed including, for instance, human, other vertebrates, and insects. Eventually, these different analyses should provide new insights into the domains and proteins contributing to microbes-mucosa and microbe-host interactions. In addition to the benefit for the identification of mucosa-specific protein features, the detailed microbe-habitat annotations could be integrated with genome sequence data and protein annotation information to provide a greater insights into the set of protein features required by microorganisms to survive in any other particular habitats.

It would be also beneficial to investigate the use of text mining to enrich the list of proteins of known subcellular location. The text-mining approach can be used to target available literature containing the experimental information of proteins and their possible locations. This should result in a limited amount of literature for manual curation. As a result, curated protein sequence data sources such as ePSORTdb could be enhanced or complemented with a greater number of proteins with known localisation, particularly for eukaryotic proteins. The outcome of this data mining process will allow the improvement of *in silico* protein subcellular localisation prediction.

In this study, only the proteomes from the 75 known mucosal-thriving microbes of Bacterial superkingdom and Protists were used for the protein clustering analysis. However, it would be interesting to also include Archaea and Fungi that are known to thrive in mucosal environments into the protein family construction. Extracytoplasmic protein sequences of a wider range of microbial taxa could be clustered based on sequence similarities in order to examine the habitat distribution in each cluster. It would then be possible to identify clusters specific to a microorganism's ecological niche, for example, the human gut or other host mucosa environments.

Finally, the work presented in this thesis is being used as a basis of two other projects. Firstly, a wet-lab experiment is planned to demonstrate the expression of M60-like-containing proteins from the known intestinal commensal, *B. thetaiotamicron* and the urogenital tract parasite, *T. vaginalis*. The experiment will also investigate the cellular localisation of the proteins in both organisms. The predicted function of the proteins as proteases and how CBMs contribute to the functions of proteins possessing M60-like domains will also be investigated. In particular, whether the CBM targets host complex glycans in a mucosal environment such as mucins. Secondly, the high-throughput sequence analysis workflow will be extended for a specific use-case that aims to, for example, find drug targets. This work is in collaboration with Glaxo-Smith-Klein and will extend the analysis workflow with particular gut microorganisms of interest.

Appendix A

List of microorganisms for which their genome were included in the project

List of 867 taxa and their taxonomic phyla or classes, abbreviation, and proteome size. The taxonomic classification was obtained from the GOLD database [1] (downloaded 22nd October 2009).

Taxonid	Organism name	Classification	Short name	Number of protein-coding genes
204669	*Acidobacteria bacterium* Ellin345	ACIDOBACTERIA	acbac	4777
234267	*Solibacter usitatus* Ellin6076	ACIDOBACTERIA	sousi	7826
351607	*Acidothermus cellulolyticus* 11B	ACTINOBACTERIA	accel	2157
290340	*Arthrobacter aurescens* TC1	ACTINOBACTERIA	araur	4587
452863	*Arthrobacter chlorophenolicus* A6	ACTINOBACTERIA	archl	4590
290399	*Arthrobacter* sp. FB24	ACTINOBACTERIA	arsp	4506
367928	*Bifidobacterium adolescentis ATCC* 15703	ACTINOBACTERIA	biado	1631
442563	*Bifidobacterium animalis* subsp. lactis AD011	ACTINOBACTERIA	biani	1528
216816	*Bifidobacterium* longum	ACTINOBACTERIA	bilon	13
205913	*Bifidobacterium longum* DJO10A	ACTINOBACTERIA	bilon	1990
206672	*Bifidobacterium longum* NCC2705	ACTINOBACTERIA	bilon	1729
391904	*Bifidobacterium longum* subsp. infantis ATCC 15697	ACTINOBACTERIA	bilon	2416
443906	*Clavibacter michiganensis* subsp. michiganensis NCPPB 382	ACTINOBACTERIA	clmic	3079
31964	*Clavibacter michiganensis* subsp. sepedonicus	ACTINOBACTERIA	clmic	3117
257309	*Corynebacterium diphtheriae NCTC* 13129	ACTINOBACTERIA	codip	2272
196164	*Corynebacterium efficiens* YS-314	ACTINOBACTERIA	coeff	2938
196627	*Corynebacterium glutamicum ATCC* 13032	ACTINOBACTERIA	coglu	6050
340322	*Corynebacterium glutamicum* R	ACTINOBACTERIA	coglu	3080
306537	*Corynebacterium jeikeium* K411	ACTINOBACTERIA	cojei	2120
504474	*Corynebacterium urealyticum DSM* 7109	ACTINOBACTERIA	coure	2024
326424	*Frankia alni* ACN14a	ACTINOBACTERIA	fraln	6711
106370	*Frankia* sp. CcI3	ACTINOBACTERIA	frsp	4499
298653	*Frankia* sp. EAN1pec	ACTINOBACTERIA	frsp	7191
266940	*Kineococcus radiotolerans* SRS30216	ACTINOBACTERIA	kirad	4681
378753	*Kocuria rhizophila* DC2201	ACTINOBACTERIA	korhi	2357
281090	*Leifsonia xyli* subsp. xyli str. CTCB07	ACTINOBACTERIA	lexyl	2030
36809	*Mycobacterium* abscessus	ACTINOBACTERIA	myabs	4941
243243	*Mycobacterium avium* 104	ACTINOBACTERIA	myavi	5120

[1] http://www.genomesonline.org/, accessed 20th August 2010

Taxonid	Organism name	Classification	Short name	Number of protein-coding genes
262316	*Mycobacterium avium* subsp. paratuberculosis K-10	ACTINOBACTERIA	myavi	4350
233413	*Mycobacterium bovis* AF2122/97	ACTINOBACTERIA	mybov	3920
410289	*Mycobacterium bovis BCG* str. Pasteur 1173P2	ACTINOBACTERIA	mybov	3952
350054	*Mycobacterium gilvum* PYR-GCK	ACTINOBACTERIA	mygil	5579
561304	*Mycobacterium leprae* Br4923	ACTINOBACTERIA	mylep	1604
272631	*Mycobacterium leprae* TN	ACTINOBACTERIA	mylep	1605
216594	*Mycobacterium marinum* M	ACTINOBACTERIA	mymar	5452
246196	*Mycobacterium smegmatis* str. MC2 155	ACTINOBACTERIA	mysme	6716
164757	*Mycobacterium* sp. JLS	ACTINOBACTERIA	mysp	5739
189918	*Mycobacterium* sp. KMS	ACTINOBACTERIA	mysp	5975
164756	*Mycobacterium* sp. MCS	ACTINOBACTERIA	mysp	5615
83331	*Mycobacterium tuberculosis* CDC1551	ACTINOBACTERIA	mytub	4189
336982	*Mycobacterium tuberculosis* F11	ACTINOBACTERIA	mytub	3941
419947	*Mycobacterium tuberculosis* H37Ra	ACTINOBACTERIA	mytub	4034
83332	*Mycobacterium tuberculosis* H37Rv	ACTINOBACTERIA	mytub	3989
362242	*Mycobacterium ulcerans* Agy99	ACTINOBACTERIA	myulc	4241
350058	*Mycobacterium vanbaalenii* PYR-1	ACTINOBACTERIA	myvan	5979
247156	*Nocardia farcinica IFM* 10152	ACTINOBACTERIA	nofar	5936
196162	*Nocardioides* sp. JS614	ACTINOBACTERIA	nosp	4909
267747	*Propionibacterium acnes* KPA171202	ACTINOBACTERIA	pracn	2297
288705	*Renibacterium salmoninarum ATCC* 33209	ACTINOBACTERIA	resal	3507
101510	*Rhodococcus jostii* RHA1	ACTINOBACTERIA	rhjos	9145
266117	*Rubrobacter xylanophilus DSM* 9941	ACTINOBACTERIA	ruxyl	3140
417289	*Saccharopolyspora erythraea NRRL* 2338	ACTINOBACTERIA	saery	7197
391037	*Salinispora arenicola* CNS-205	ACTINOBACTERIA	saare	4917
369723	*Salinispora tropica* CNB-440	ACTINOBACTERIA	satro	4536
227882	*Streptomyces avermitilis* MA-4680	ACTINOBACTERIA	stave	7676
100226	*Streptomyces coelicolor* A3(2)	ACTINOBACTERIA	stcoe	8154
455632	*Streptomyces griseus* subsp. griseus NBRC 13350	ACTINOBACTERIA	stgri	7136
269800	*Thermobifida fusca* YX	ACTINOBACTERIA	thfus	3110
203267	*Tropheryma whipplei* str. Twist	ACTINOBACTERIA	trwhi	808
218496	*Tropheryma whipplei* TW08/27	ACTINOBACTERIA	trwhi	783
224324	*Aquifex aeolicus* VF5	AQUIFICAE	aqaeo	1560
380749	*Hydrogenobaculum* sp. Y04AAS1	AQUIFICAE	hysp	1629
436114	*Sulfurihydrogenibium* sp. YO3AOP1	AQUIFICAE	susp	1721
272559	*Bacteroides fragilis NCTC* 9343	BACTEROIDETES	bafra	4231
295405	*Bacteroides fragilis* YCH46	BACTEROIDETES	bafra	4625
226186	*Bacteroides thetaiotaomicron* VPI-5482	BACTEROIDETES	bathe	4816
435590	*Bacteroides vulgatus ATCC* 8482	BACTEROIDETES	bavul	4065
452471	*Candidatus Amoebophilus asiaticus* 5a2	BACTEROIDETES	caamo	1283
511995	*Candidatus Azobacteroides pseudotrichonymphae genomovar.* CFP2	BACTEROIDETES	caazo	852
444179	*Candidatus Sulcia muelleri* GWSS	BACTEROIDETES	casul	227
269798	*Cytophaga hutchinsonii ATCC* 33406	BACTEROIDETES	cyhut	3785
376686	*Flavobacterium johnsoniae* UW101	BACTEROIDETES	fljoh	5017
402612	*Flavobacterium psychrophilum* JIP02/86	BACTEROIDETES	flpsy	2412
411154	*Gramella forsetii* KT0803	BACTEROIDETES	grfor	3584
435591	*Parabacteroides distasonis ATCC* 8503	BACTEROIDETES	padis	3850
431947	*Porphyromonas gingivalis ATCC* 33277	BACTEROIDETES	pogin	2090
242619	*Porphyromonas gingivalis* W83	BACTEROIDETES	pogin	1909
309807	*Salinibacter ruber DSM* 13855	BACTEROIDETES	sarub	2833
264201	*Candidatus Protochlamydia amoebophila* UWE25	CHLAMYDIAE	capro	2031
243161	*Chlamydia muridarum* Nigg	CHLAMYDIAE	chmur	911
471472	*Chlamydia trachomatis* 434/Bu	CHLAMYDIAE	chtra	874
315277	*Chlamydia trachomatis* A/HAR-13	CHLAMYDIAE	chtra	919
272561	*Chlamydia trachomatis* D/UW-3/CX	CHLAMYDIAE	chtra	895
471473	*Chlamydia trachomatis* L2b/UCH-1/proctitis	CHLAMYDIAE	chtra	874
218497	*Chlamydophila abortus* S26/3	CHLAMYDIAE	chabo	932
227941	*Chlamydophila caviae* GPIC	CHLAMYDIAE	chcav	1005
264202	*Chlamydophila felis* Fe/C-56	CHLAMYDIAE	chfel	1013
115711	*Chlamydophila pneumoniae* AR39	CHLAMYDIAE	chpne	1112
115713	*Chlamydophila pneumoniae* CWL029	CHLAMYDIAE	chpne	1052
138677	*Chlamydophila pneumoniae* J138	CHLAMYDIAE	chpne	1069
182082	*Chlamydophila pneumoniae* TW-183	CHLAMYDIAE	chpne	1113

Taxonid	Organism name	Classification	Short name	Number of protein-coding genes
517417	*Chlorobaculum parvum NCIB* 8327	CHLOROBI	chpar	2043
340177	*Chlorobium chlorochromatii* CaD3	CHLOROBI	chchl	2002
290315	*Chlorobium limicola DSM* 245	CHLOROBI	chlim	2434
331678	*Chlorobium phaeobacteroides* BS1	CHLOROBI	chpha	2469
290317	*Chlorobium phaeobacteroides DSM* 266	CHLOROBI	chpha	2650
290318	*Chlorobium phaeovibrioides DSM 265 (Prosthecochloris vibrioformis DSM* 265)	CHLOROBI	chpha	1753
194439	*Chlorobium tepidum* TLS	CHLOROBI	chtep	2245
517418	*Chloroherpeton thalassium ATCC* 35110	CHLOROBI	chtha	2710
319225	*Pelodictyon luteolum DSM* 273	CHLOROBI	pelut	2083
324925	*Pelodictyon phaeoclathratiforme* BU-1	CHLOROBI	pepha	2707
290512	*Prosthecochloris aestuarii DSM* 271	CHLOROBI	praes	2327
326427	*Chloroflexus aggregans DSM* 9485	CHLOROFLEXI	chagg	3730
324602	*Chloroflexus aurantiacus* J-10-fl	CHLOROFLEXI	chaur	3853
480224	*Chloroflexus* sp. Y-400-fl	CHLOROFLEXI	chsp	4159
243164	*Dehalococcoides ethenogenes* 195	CHLOROFLEXI	deeth	1580
216389	*Dehalococcoides* sp. BAV1	CHLOROFLEXI	desp	1371
255470	*Dehalococcoides* sp. CBDB1	CHLOROFLEXI	desp	1458
316274	*Herpetosiphon aurantiacus ATCC* 23779	CHLOROFLEXI	heaur	5278
383372	*Roseiflexus castenholzii DSM* 13941	CHLOROFLEXI	rocas	4330
357808	*Roseiflexus* sp. RS-1	CHLOROFLEXI	rosp	4517
309801	*Thermomicrobium roseum DSM* 5159	CHLOROFLEXI	thros	2854
436308	*Nitrosopumilus maritimus* SCM1	CRENARCHAEOTA-CENARCHAEA	nimar	1795
272557	*Aeropyrum pernix* K1	CRENARCHAEOTA-THERMOPROTEI	aeper	1700
397948	*Caldivirga maquilingensis* IC-167	CRENARCHAEOTA-THERMOPROTEI	camaq	1963
490899	*Desulfurococcus kamchatkensis* 1221n	CRENARCHAEOTA-THERMOPROTEI	dekam	1471
415426	*Hyperthermus butylicus DSM* 5456	CRENARCHAEOTA-THERMOPROTEI	hybut	1602
453591	*Ignicoccus hospitalis* KIN4/I	CRENARCHAEOTA-THERMOPROTEI	ighos	1434
399549	*Metallosphaera sedula DSM* 5348	CRENARCHAEOTA-THERMOPROTEI	mesed	2256
178306	*Pyrobaculum aerophilum* str. IM2	CRENARCHAEOTA-THERMOPROTEI	pyaer	2605
340102	*Pyrobaculum arsenaticum DSM* 13514	CRENARCHAEOTA-THERMOPROTEI	pyars	2299
410359	*Pyrobaculum calidifontis JCM* 11548	CRENARCHAEOTA-THERMOPROTEI	pycal	2149
384616	*Pyrobaculum islandicum DSM* 4184	CRENARCHAEOTA-THERMOPROTEI	pyisl	1978
399550	*Staphylothermus marinus* F1	CRENARCHAEOTA-THERMOPROTEI	stmar	1570
330779	*Sulfolobus acidocaldarius DSM* 639	CRENARCHAEOTA-THERMOPROTEI	suaci	2223
273057	*Sulfolobus solfataricus* P2	CRENARCHAEOTA-THERMOPROTEI	susol	2977
273063	*Sulfolobus tokodaii* str. 7	CRENARCHAEOTA-THERMOPROTEI	sutok	2825
368408	*Thermofilum pendens Hrk* 5	CRENARCHAEOTA-THERMOPROTEI	thpen	1876
444157	*Thermoproteus neutrophilus* V24Sta	CRENARCHAEOTA-THERMOPROTEI	thneu	1966
329726	*Acaryochloris marina* MBIC11017	CYANOBACTERIA	acmar	8383
240292	*Anabaena variabilis ATCC 29413 (Anabaena flos-aquae UTEX* 1444)	CYANOBACTERIA	anvar	5661
43989	*Cyanothece* sp. ATCC 51142	CYANOBACTERIA	cysp	5304
65393	*Cyanothece* sp. PCC 7424	CYANOBACTERIA	cysp	5710
395961	*Cyanothece* sp. PCC 7425	CYANOBACTERIA	cysp	5327
41431	*Cyanothece* sp. PCC 8801	CYANOBACTERIA	cysp	4367
251221	*Gloeobacter violaceus PCC* 7421	CYANOBACTERIA	glvio	4430
449447	*Microcystis aeruginosa* NIES-843	CYANOBACTERIA	miaer	6312
63737	*Nostoc punctiforme PCC 73102 (Nostoc punctiforme ATCC* 29133)	CYANOBACTERIA	nopun	6690

Taxonid	Organism name	Classification	Short name	Number of protein-coding genes
103690	*Nostoc* sp. PCC 7120 (Anabaena sp. PCC7120)	CYANOBACTERIA	nosp	6130
146891	*Prochlorococcus marinus* str. AS9601	CYANOBACTERIA	prmar	1921
93059	*Prochlorococcus marinus* str. MIT 9211	CYANOBACTERIA	prmar	1855
93060	*Prochlorococcus marinus* str. MIT 9215	CYANOBACTERIA	prmar	1983
167546	*Prochlorococcus marinus* str. MIT 9301	CYANOBACTERIA	prmar	1907
59922	*Prochlorococcus marinus* str. MIT 9303	CYANOBACTERIA	prmar	2997
74546	*Prochlorococcus marinus* str. MIT 9312	CYANOBACTERIA	prmar	1810
74547	*Prochlorococcus marinus* str. MIT 9313	CYANOBACTERIA	prmar	2269
167542	*Prochlorococcus marinus* str. MIT 9515	CYANOBACTERIA	prmar	1906
167555	*Prochlorococcus marinus* str. NATL1A	CYANOBACTERIA	prmar	2193
59920	*Prochlorococcus marinus* str. NATL2A	CYANOBACTERIA	prmar	2163
167539	*Prochlorococcus marinus* subsp. marinus str. CCMP1375 (Prochlorococcus marinus SS120)	CYANOBACTERIA	prmar	1883
59919	*Prochlorococcus marinus* subsp. pastoris str. CCMP1986 (Prochlorococcus marinus MED4)	CYANOBACTERIA	prmar	1717
269084	*Synechococcus elongatus PCC* 6301	CYANOBACTERIA	syelo	2527
1140	*Synechococcus elongatus PCC* 7942	CYANOBACTERIA	syelo	2662
64471	*Synechococcus* sp. CC9311	CYANOBACTERIA	sysp	2892
110662	*Synechococcus* sp. CC9605	CYANOBACTERIA	sysp	2645
316279	*Synechococcus* sp. CC9902	CYANOBACTERIA	sysp	2307
321332	*Synechococcus* sp. JA-2-3B'a(2-13)	CYANOBACTERIA	sysp	2862
32049	*Synechococcus* sp. PCC 7002	CYANOBACTERIA	sysp	3186
316278	*Synechococcus* sp. RCC307	CYANOBACTERIA	sysp	2535
32051	*Synechococcus* sp. WH 7803	CYANOBACTERIA	sysp	2533
84588	*Synechococcus* sp. WH 8102	CYANOBACTERIA	sysp	2519
1148	*Synechocystis* sp. PCC 6803 (Synechocystis PCC6803)	CYANOBACTERIA	sysp	3569
197221	*Thermosynechococcus elongatus* BP-1	CYANOBACTERIA	thelo	2476
203124	*Trichodesmium erythraeum* IMS101	CYANOBACTERIA	trery	4451
309799	*Dictyoglomus thermophilum* H-6-12	DICTYOGLOMI	dithe	1912
515635	*Dictyoglomus turgidum DSM* 6724	DICTYOGLOMI	ditur	1744
445932	*Elusimicrobium minutum* Pei191	ELUSIMICROBIA	elmin	1529
521011	*Candidatus Methanosphaerula palustris* E1-9c	EURYARCHAEOTA	camet	2655
224325	*Archaeoglobus fulgidus DSM* 4304	EURYARCHAEOTA-ARCHAEOGLOBI	arful	2420
272569	*Haloarcula marismortui ATCC* 43049	EURYARCHAEOTA-HALOBACTERIA	hamar	4240
478009	*Halobacterium salinarum* R1	EURYARCHAEOTA-HALOBACTERIA	hasal	2749
64091	*Halobacterium* sp. NRC-1	EURYARCHAEOTA-HALOBACTERIA	hasp	2622
362976	*Haloquadratum walsbyi DSM* 16790	EURYARCHAEOTA-HALOBACTERIA	hawal	2646
416348	*Halorubrum lacusprofundi ATCC* 49239	EURYARCHAEOTA-HALOBACTERIA	halac	3560
348780	*Natronomonas pharaonis DSM* 2160	EURYARCHAEOTA-HALOBACTERIA	napha	2822
420247	*Methanobrevibacter smithii ATCC* 35061	EURYARCHAEOTA-METHANOBACTERIA	mesmi	1793
339860	*Methanosphaera stadtmanae DSM* 3091	EURYARCHAEOTA-METHANOBACTERIA	mesta	1534
187420	*Methanothermobacter thermautotrophicus* str. Delta H (Methanobacterium thermoautotrophicum str. deltaH)	EURYARCHAEOTA-METHANOBACTERIA	methe	1873
243232	*Methanocaldococcus jannaschii DSM* 2661	EURYARCHAEOTA-METHANOCOCCI	mejan	1786
419665	*Methanococcus aeolicus* Nankai-3	EURYARCHAEOTA-METHANOCOCCI	meaeo	1490
402880	*Methanococcus maripaludis* C5	EURYARCHAEOTA-METHANOCOCCI	memar	1822
444158	*Methanococcus maripaludis* C6	EURYARCHAEOTA-METHANOCOCCI	memar	1826
426368	*Methanococcus maripaludis* C7	EURYARCHAEOTA-METHANOCOCCI	memar	1788
267377	*Methanococcus maripaludis* S2	EURYARCHAEOTA-METHANOCOCCI	memar	1722

Taxonid	Organism name	Classification	Short name	Number of protein-coding genes
406327	*Methanococcus vannielii* SB	EURYARCHAEOTA-METHANOCOCCI	mevan	1678
456442	*Candidatus Methanoregula boonei* 6A8	EURYARCHAEOTA-METHANOMICROBIA	camet	2450
259564	*Methanococcoides burtonii DSM* 6242	EURYARCHAEOTA-METHANOMICROBIA	mebur	2273
410358	*Methanocorpusculum labreanum* Z	EURYARCHAEOTA-METHANOMICROBIA	melab	1739
368407	*Methanoculleus marisnigri* JR1	EURYARCHAEOTA-METHANOMICROBIA	memar	2489
349307	*Methanosaeta thermophila PT (Methanothrix thermophila* PT)	EURYARCHAEOTA-METHANOMICROBIA	methe	1696
188937	*Methanosarcina acetivorans* C2A	EURYARCHAEOTA-METHANOMICROBIA	meace	4540
269797	*Methanosarcina barkeri* str. Fusaro	EURYARCHAEOTA-METHANOMICROBIA	mebar	3624
192952	*Methanosarcina mazei* Go1	EURYARCHAEOTA-METHANOMICROBIA	memaz	3370
323259	*Methanospirillum hungatei* JF-1	EURYARCHAEOTA-METHANOMICROBIA	mehun	3139
190192	*Methanopyrus kandleri* AV19	EURYARCHAEOTA-METHANOPYRI	mekan	1687
272844	*Pyrococcus abyssi* GE5	EURYARCHAEOTA-THERMOCOCCI	pyaby	1782
186497	*Pyrococcus furiosus DSM* 3638	EURYARCHAEOTA-THERMOCOCCI	pyfur	2125
70601	*Pyrococcus horikoshii* OT3	EURYARCHAEOTA-THERMOCOCCI	pyhor	1955
69014	*Thermococcus kodakarensis* KOD1	EURYARCHAEOTA-THERMOCOCCI	thkod	2306
523850	*Thermococcus onnurineus* NA1	EURYARCHAEOTA-THERMOCOCCI	thonn	1976
263820	*Picrophilus torridus DSM* 9790	EURYARCHAEOTA-THERMOPLASMATA	pitor	1535
273075	*Thermoplasma acidophilum DSM* 1728	EURYARCHAEOTA-THERMOPLASMATA	thaci	1482
273116	*Thermoplasma volcanium* GSS1	EURYARCHAEOTA-THERMOPLASMATA	thvol	1499
441768	*Acholeplasma laidlawii* PG-8A	FIRMICUTES	aclai	1380
293826	*Alkaliphilus metalliredigens* QYMF	FIRMICUTES	almet	4625
350688	*Alkaliphilus oremlandii* OhILAs	FIRMICUTES	alore	2836
521460	*Anaerocellum thermophilum DSM* 6725	FIRMICUTES	anthe	2666
491915	*Anoxybacillus flavithermus* WK1	FIRMICUTES	anfla	2832
322098	Aster yellows witches'-broom phytoplasma AYWB	FIRMICUTES	asyel	693
326423	*Bacillus amyloliquefaciens* FZB42	FIRMICUTES	baamy	3693
261594	*Bacillus anthracis* str. 'Ames Ancestor'	FIRMICUTES	baant	5584
198094	*Bacillus anthracis* str. Ames	FIRMICUTES	baant	5311
260799	*Bacillus anthracis* str. Sterne	FIRMICUTES	baant	5287
405534	*Bacillus cereus* AH187	FIRMICUTES	bacer	5758
405535	*Bacillus cereus* AH820	FIRMICUTES	bacer	5810
222523	*Bacillus cereus ATCC* 10987	FIRMICUTES	bacer	5844
226900	*Bacillus cereus ATCC* 14579	FIRMICUTES	bacer	5255
405532	*Bacillus cereus* B4264	FIRMICUTES	bacer	5408
288681	*Bacillus cereus* E33L	FIRMICUTES	bacer	5641
405531	*Bacillus cereus* G9842	FIRMICUTES	bacer	5857
361100	*Bacillus cereus* Q1	FIRMICUTES	bacer	5488
315749	*Bacillus cereus* subsp. cytotoxis NVH 391-98	FIRMICUTES	bacer	3844
66692	*Bacillus clausii* KSM-K16	FIRMICUTES	bacla	4096
272558	*Bacillus halodurans* C-125	FIRMICUTES	bahal	4066
279010	*Bacillus licheniformis ATCC 14580 (DSM* 13)	FIRMICUTES	balic	4196
315750	*Bacillus pumilus* SAFR-032	FIRMICUTES	bapum	3681
224308	*Bacillus subtilis* subsp. subtilis str. 168	FIRMICUTES	basub	4105
281309	*Bacillus thuringiensis* serovar konkukian str. 97-27	FIRMICUTES	bathu	5197
412694	*Bacillus thuringiensis* str. Al Hakam	FIRMICUTES	bathu	4798
315730	*Bacillus weihenstephanensis* KBAB4	FIRMICUTES	bawei	5653

Taxonid	Organism name	Classification	Short name	Number of protein-coding genes
351627	*Caldicellulosiruptor saccharolyticus DSM* 8903	FIRMICUTES	casac	2679
477974	*Candidatus Desulforudis audaxviator* MP104C	FIRMICUTES	cades	2157
59748	*Candidatus Phytoplasma* australiense	FIRMICUTES	caphy	684
37692	*Candidatus Phytoplasma* mali	FIRMICUTES	caphy	479
246194	*Carboxydothermus hydrogenoformans* Z-2901	FIRMICUTES	cahyd	2620
272562	*Clostridium acetobutylicum ATCC* 824	FIRMICUTES	clace	3848
290402	*Clostridium beijerinckii NCIMB* 8052	FIRMICUTES	clbei	5020
441770	*Clostridium botulinum A* str. ATCC 19397	FIRMICUTES	clbot	3548
413999	*Clostridium botulinum A* str. ATCC 3502	FIRMICUTES	clbot	3590
441771	*Clostridium botulinum A* str. Hall	FIRMICUTES	clbot	3404
498214	*Clostridium botulinum A3* str. Loch Maree	FIRMICUTES	clbot	3984
508765	*Clostridium botulinum B* str. Eklund 17B	FIRMICUTES	clbot	3527
498213	*Clostridium botulinum B1* str. Okra	FIRMICUTES	clbot	3852
508767	*Clostridium botulinum E3* str. Alaska E43	FIRMICUTES	clbot	3256
441772	*Clostridium botulinum F* str. Langeland	FIRMICUTES	clbot	3659
394503	*Clostridium cellulolyticum* H10	FIRMICUTES	clcel	3390
272563	*Clostridium difficile* 630	FIRMICUTES	cldif	3753
431943	*Clostridium kluyveri DSM* 555	FIRMICUTES	clklu	3913
583346	*Clostridium kluyveri NBRC* 12016	FIRMICUTES	clklu	3523
386415	*Clostridium novyi* NT	FIRMICUTES	clnov	2315
195103	*Clostridium perfringens ATCC* 13124	FIRMICUTES	clper	2876
289380	*Clostridium perfringens* SM101	FIRMICUTES	clper	2578
195102	*Clostridium perfringens* str. 13	FIRMICUTES	clper	2723
357809	*Clostridium phytofermentans* ISDg	FIRMICUTES	clphy	3902
212717	*Clostridium tetani* E88	FIRMICUTES	cltet	2432
203119	*Clostridium thermocellum ATCC* 27405	FIRMICUTES	clthe	3189
309798	*Coprothermobacter proteolyticus DSM* 5265	FIRMICUTES	copro	1482
272564	*Desulfitobacterium hafniense* DCB-2	FIRMICUTES	dehaf	4883
138119	*Desulfitobacterium hafniense* Y51	FIRMICUTES	dehaf	5060
349161	*Desulfotomaculum reducens* MI-1	FIRMICUTES	dered	3276
226185	*Enterococcus faecalis* V583	FIRMICUTES	enfae	3265
262543	*Exiguobacterium sibiricum* 255-15	FIRMICUTES	exsib	3015
334413	*Finegoldia magna ATCC* 29328	FIRMICUTES	fimag	1813
235909	*Geobacillus kaustophilus* HTA426	FIRMICUTES	gekau	3540
420246	*Geobacillus thermodenitrificans* NG80-2	FIRMICUTES	gethe	3445
373903	*Halothermothrix orenii H* 168	FIRMICUTES	haore	2342
498761	*Heliobacterium modesticaldum* Ice1	FIRMICUTES	hemod	3000
272621	*Lactobacillus acidophilus* NCFM	FIRMICUTES	laaci	1862
387344	*Lactobacillus brevis ATCC* 367	FIRMICUTES	labre	2218
321967	*Lactobacillus casei ATCC* 334	FIRMICUTES	lacas	2771
543734	*Lactobacillus casei* BL23	FIRMICUTES	lacas	3044
390333	*Lactobacillus delbrueckii* subsp. bulgaricus ATCC 11842	FIRMICUTES	ladel	1562
321956	*Lactobacillus delbrueckii* subsp. bulgaricus ATCC BAA-365	FIRMICUTES	ladel	1721
334390	*Lactobacillus fermentum IFO* 3956	FIRMICUTES	lafer	1843
324831	*Lactobacillus gasseri ATCC* 33323	FIRMICUTES	lagas	1755
405566	*Lactobacillus helveticus DPC* 4571	FIRMICUTES	lahel	1610
257314	*Lactobacillus johnsonii NCC* 533	FIRMICUTES	lajoh	1821
220668	*Lactobacillus plantarum* WCFS1	FIRMICUTES	lapla	3057
557436	*Lactobacillus reuteri DSM* 20016	FIRMICUTES	lareu	1900
557433	*Lactobacillus reuteri JCM* 1112	FIRMICUTES	lareu	1820
314315	*Lactobacillus sakei* subsp. sakei 23K	FIRMICUTES	lasak	1879
362948	*Lactobacillus salivarius* UCC118	FIRMICUTES	lasal	296
364252	*Lactobacillus salivarius* UCC118	FIRMICUTES	lasal	1717
416870	*Lactococcus lactis* subsp. cremoris MG1363	FIRMICUTES	lalac	2434
272622	*Lactococcus lactis* subsp. cremoris SK11	FIRMICUTES	lalac	2504
272623	*Lactococcus lactis* subsp. lactis Il1403	FIRMICUTES	lalac	2321
349519	*Leuconostoc citreum* KM20	FIRMICUTES	lecit	1820
203120	*Leuconostoc mesenteroides* subsp. mesenteroides ATCC 8293	FIRMICUTES	lemes	2005
272626	*Listeria innocua* Clip11262	FIRMICUTES	liinn	3043
169963	*Listeria monocytogenes* EGD-e	FIRMICUTES	limon	2846
552536	*Listeria monocytogenes* HCC23	FIRMICUTES	limon	2974
265669	*Listeria monocytogenes* str. 4b F2365	FIRMICUTES	limon	2821
386043	*Listeria welshimeri* serovar 6b str. SLCC5334	FIRMICUTES	liwel	2774
444177	*Lysinibacillus sphaericus* C3-41	FIRMICUTES	lysph	4771

Taxonid	Organism name	Classification	Short name	Number of protein-coding genes
458233	*Macrococcus caseolyticus* JCSC5402	FIRMICUTES	macas	2052
265311	*Mesoplasma florum* L1	FIRMICUTES	meflo	682
264732	*Moorella thermoacetica ATCC* 39073	FIRMICUTES	mothe	2463
457570	*Natranaerobius thermophilus* JW/NM-WN-LF	FIRMICUTES	nathe	2906
221109	*Oceanobacillus iheyensis* HTE831	FIRMICUTES	ocihe	3500
203123	*Oenococcus oeni* PSU-1	FIRMICUTES	oeoen	1691
262768	*Onion yellows phytoplasma* OY-M	FIRMICUTES	onyel	754
278197	*Pediococcus pentosaceus ATCC* 25745	FIRMICUTES	pepen	1755
370438	*Pelotomaculum thermopropionicum* SI	FIRMICUTES	pethe	2920
273036	*Staphylococcus aureus* RF122	FIRMICUTES	staur	2509
93062	*Staphylococcus aureus* subsp. aureus COL	FIRMICUTES	staur	2615
359787	*Staphylococcus aureus* subsp. aureus JH1	FIRMICUTES	staur	2780
359786	*Staphylococcus aureus* subsp. aureus JH9	FIRMICUTES	staur	2726
282458	*Staphylococcus aureus* subsp. aureus MRSA252	FIRMICUTES	staur	2656
282459	*Staphylococcus aureus* subsp. aureus MSSA476	FIRMICUTES	staur	2598
418127	*Staphylococcus aureus* subsp. aureus Mu3	FIRMICUTES	staur	2698
158878	*Staphylococcus aureus* subsp. aureus Mu50	FIRMICUTES	staur	2731
196620	*Staphylococcus aureus* subsp. aureus MW2	FIRMICUTES	staur	2632
158879	*Staphylococcus aureus* subsp. aureus N315	FIRMICUTES	staur	2619
93061	*Staphylococcus aureus* subsp. aureus NCTC 8325	FIRMICUTES	staur	2892
426430	*Staphylococcus aureus* subsp. aureus str. Newman	FIRMICUTES	staur	2614
367830	*Staphylococcus aureus* subsp. aureus USA300	FIRMICUTES	staur	2604
451516	*Staphylococcus aureus* subsp. aureus USA300_TCH1516	FIRMICUTES	staur	2683
176280	*Staphylococcus epidermidis ATCC* 12228	FIRMICUTES	stepi	2485
176279	*Staphylococcus epidermidis* RP62A	FIRMICUTES	stepi	2526
279808	*Staphylococcus haemolyticus* JCSC1435	FIRMICUTES	sthae	2692
342451	*Staphylococcus saprophyticus* subsp. saprophyticus ATCC 15305	FIRMICUTES	stsap	2514
208435	*Streptococcus agalactiae* 2603V/R	FIRMICUTES	staga	2124
205921	*Streptococcus agalactiae* A909	FIRMICUTES	staga	1996
211110	*Streptococcus agalactiae* NEM316	FIRMICUTES	staga	2094
552526	*Streptococcus equi* subsp. zooepidemicus MGCS10565	FIRMICUTES	stequ	1893
467705	*Streptococcus gordonii* str. Challis substr. CH1	FIRMICUTES	stgor	2051
210007	*Streptococcus mutans* UA159	FIRMICUTES	stmut	1960
561276	*Streptococcus pneumoniae ATCC* 700669	FIRMICUTES	stpne	1990
516950	*Streptococcus pneumoniae* CGSP14	FIRMICUTES	stpne	2206
373153	*Streptococcus pneumoniae* D39	FIRMICUTES	stpne	1914
512566	*Streptococcus pneumoniae* G54	FIRMICUTES	stpne	2115
487214	*Streptococcus pneumoniae* Hungary19A-6	FIRMICUTES	stpne	2155
171101	*Streptococcus pneumoniae* R6	FIRMICUTES	stpne	2043
170187	*Streptococcus pneumoniae* TIGR4	FIRMICUTES	stpne	2105
430513	*Streptococcus pyogenes M1* GAS	FIRMICUTES	stpyo	1697
370559	*Streptococcus pyogenes* MGAS10270	FIRMICUTES	stpyo	1986
286636	*Streptococcus pyogenes* MGAS10394	FIRMICUTES	stpyo	1886
370568	*Streptococcus pyogenes* MGAS10750	FIRMICUTES	stpyo	1979
370553	*Streptococcus pyogenes* MGAS2096	FIRMICUTES	stpyo	1898
198466	*Streptococcus pyogenes* MGAS315	FIRMICUTES	stpyo	1865
293713	*Streptococcus pyogenes* MGAS5005	FIRMICUTES	stpyo	1865
319710	*Streptococcus pyogenes* MGAS6180	FIRMICUTES	stpyo	1894
186103	*Streptococcus pyogenes* MGAS8232	FIRMICUTES	stpyo	1839
370558	*Streptococcus pyogenes* MGAS9429	FIRMICUTES	stpyo	1877
471876	*Streptococcus pyogenes* NZ131	FIRMICUTES	stpyo	1700
193567	*Streptococcus pyogenes* SSI-1	FIRMICUTES	stpyo	1861
160491	*Streptococcus pyogenes* str. Manfredo	FIRMICUTES	stpyo	1745
388919	*Streptococcus sanguinis* SK36	FIRMICUTES	stsan	2270
391295	*Streptococcus suis* 05ZYH33	FIRMICUTES	stsui	2186
391296	*Streptococcus suis* 98HAH33	FIRMICUTES	stsui	2185
299768	*Streptococcus thermophilus* CNRZ1066	FIRMICUTES	stthe	1915
322159	*Streptococcus thermophilus* LMD-9	FIRMICUTES	stthe	1716
264199	*Streptococcus thermophilus LMG* 18311	FIRMICUTES	stthe	1889
218495	*Streptococcus uberis* 0140J	FIRMICUTES	stube	1760
292459	*Symbiobacterium thermophilum IAM* 14863	FIRMICUTES	sythe	3338

Taxonid	Organism name	Classification	Short name	Number of protein-coding genes
335541	*Syntrophomonas wolfei* subsp. wolfei str. Goettingen	FIRMICUTES	sywol	2504
340099	*Thermoanaerobacter pseudethanolicus ATCC* 33223	FIRMICUTES	thpse	2243
399726	*Thermoanaerobacter* sp. X514	FIRMICUTES	thsp	2349
273068	*Thermoanaerobacter tengcongensis* MB4	FIRMICUTES	thten	2588
505682	*Ureaplasma parvum* serovar 3 str. ATCC 27815	FIRMICUTES	urpar	609
273119	*Ureaplasma parvum* serovar 3 str. ATCC 700970	FIRMICUTES	urpar	614
565575	*Ureaplasma urealyticum* serovar 10 str. ATCC 33699	FIRMICUTES	urure	646
344612	*Aspergillus clavatus NRRL* 1	FUNGI-ASCOMYCOTA	ascla	9121
330879	*Aspergillus fumigatus* Af293	FUNGI-ASCOMYCOTA	asfum	9630
227321	*Aspergillus nidulans FGSC* A4	FUNGI-ASCOMYCOTA	asnid	9410
341663	*Aspergillus terreus* NIH2624	FUNGI-ASCOMYCOTA	aster	10406
237561	*Candida albicans* SC5314	FUNGI-ASCOMYCOTA	caalb	403
284593	*Candida glabrata CBS* 138	FUNGI-ASCOMYCOTA	cagla	5192
284592	*Debaryomyces hansenii* CBS767	FUNGI-ASCOMYCOTA	dehan	6316
229533	*Gibberella zeae PH-1 (anamorph: Fusarium* graminearum)	FUNGI-ASCOMYCOTA	gizea	11578
284590	*Kluyveromyces lactis NRRL* Y-1140	FUNGI-ASCOMYCOTA	kllac	5327
242507	*Magnaporthe grisea* 70-15	FUNGI-ASCOMYCOTA	magri	1178
367110	*Neurospora crassa* OR74A	FUNGI-ASCOMYCOTA	necra	10079
322104	*Pichia stipitis CBS* 6054	FUNGI-ASCOMYCOTA	pisti	5816
4932	*Saccharomyces cerevisiae (baker's* yeast)	FUNGI-ASCOMYCOTA	sacer	5880
4896	*Schizosaccharomyces pombe (fission* yeast)	FUNGI-ASCOMYCOTA	scpom	5003
284591	*Yarrowia lipolytica* CLIB122	FUNGI-ASCOMYCOTA	yalip	6448
283643	*Cryptococcus neoformans* var. neoformans B-3501A	FUNGI-BASIDIOMYCOTA	crneo	6578
214684	*Cryptococcus neoformans* var. neoformans JEC21 (Filobasidiella neoformans var. neoformans strain JEC21)	FUNGI-BASIDIOMYCOTA	crneo	6475
237631	*Ustilago maydis* 521	FUNGI-BASIDIOMYCOTA	usmay	6522
284813	*Encephalitozoon cuniculi* GB-M1	FUNGI-MICROSPORIDIA	encun	1996
190304	*Fusobacterium nucleatum* subsp. nucleatum ATCC 25586	FUSOBACTERIA	funuc	2067
374847	*Candidatus Korarchaeum cryptofilum* OPF8	KORARCHAEOTA	cakor	1602
228908	*Nanoarchaeum equitans* Kin4-M	NANOARCHAEOTA	naequ	536
289376	*Thermodesulfovibrio yellowstonii DSM* 11347	NITROSPIRAE	thyel	2033
243090	*Rhodopirellula baltica SH* 1	PLANCTOMYCETES	rhbal	7325
349163	*Acidiphilium cryptum* JF-5	PROTEOBACTERIA-ALPHA	accry	3559
311403	*Agrobacterium radiobacter* K84	PROTEOBACTERIA-ALPHA	agrad	6684
176299	*Agrobacterium tumefaciens* str. C58	PROTEOBACTERIA-ALPHA	agtum	5355
311402	*Agrobacterium vitis* S4	PROTEOBACTERIA-ALPHA	agvit	5389
320483	*Anaplasma marginale* str. Florida	PROTEOBACTERIA-ALPHA	anmar	940
234826	*Anaplasma marginale* str. St. Maries	PROTEOBACTERIA-ALPHA	anmar	948
212042	*Anaplasma phagocytophilum* HZ	PROTEOBACTERIA-ALPHA	anpha	1264
438753	*Azorhizobium caulinodans ORS* 571	PROTEOBACTERIA-ALPHA	azcau	4717
360095	*Bartonella bacilliformis* KC583	PROTEOBACTERIA-ALPHA	babac	1283
283166	*Bartonella henselae* str. Houston-1	PROTEOBACTERIA-ALPHA	bahen	1488
283165	*Bartonella quintana* str. Toulouse	PROTEOBACTERIA-ALPHA	baqui	1142
382640	*Bartonella tribocorum CIP* 105476	PROTEOBACTERIA-ALPHA	batri	2092
395963	*Beijerinckia indica* subsp. indica ATCC 9039	PROTEOBACTERIA-ALPHA	beind	3784
224911	*Bradyrhizobium japonicum USDA* 110	PROTEOBACTERIA-ALPHA	brjap	8317
288000	*Bradyrhizobium* sp. BTAi1	PROTEOBACTERIA-ALPHA	brsp	7622
114615	*Bradyrhizobium* sp. ORS278	PROTEOBACTERIA-ALPHA	brsp	6717
262698	*Brucella abortus bv. 1* str. 9-941	PROTEOBACTERIA-ALPHA	brabo	3085
430066	*Brucella abortus* S19	PROTEOBACTERIA-ALPHA	brabo	3000
483179	*Brucella canis ATCC* 23365	PROTEOBACTERIA-ALPHA	brcan	3251
224914	*Brucella melitensis* 16M	PROTEOBACTERIA-ALPHA	brmel	3198
359391	*Brucella melitensis biovar Abortus* 2308	PROTEOBACTERIA-ALPHA	brmel	3034
444178	*Brucella ovis ATCC* 25840	PROTEOBACTERIA-ALPHA	brovi	2890
204722	*Brucella suis* 1330	PROTEOBACTERIA-ALPHA	brsui	3272
470137	*Brucella suis ATCC* 23445	PROTEOBACTERIA-ALPHA	brsui	3241
335992	*Candidatus Pelagibacter ubique* HTCC1062	PROTEOBACTERIA-ALPHA	capel	1354
190650	*Caulobacter crescentus* CB15	PROTEOBACTERIA-ALPHA	cacre	3737
565050	*Caulobacter crescentus* NA1000	PROTEOBACTERIA-ALPHA	cacre	3876

Taxonid	Organism name	Classification	Short name	Number of protein-coding genes
366602	*Caulobacter* sp. K31	PROTEOBACTERIA-ALPHA	casp	5438
398580	*Dinoroseobacter shibae DFL* 12	PROTEOBACTERIA-ALPHA	dishi	4187
269484	*Ehrlichia canis* str. Jake	PROTEOBACTERIA-ALPHA	ehcan	925
205920	*Ehrlichia chaffeensis* str. Arkansas	PROTEOBACTERIA-ALPHA	ehcha	1105
302409	*Ehrlichia ruminantium* str. Gardel	PROTEOBACTERIA-ALPHA	ehrum	950
254945	*Ehrlichia ruminantium* str. Welgevonden	PROTEOBACTERIA-ALPHA	ehrum	1846
314225	*Erythrobacter litoralis* HTCC2594	PROTEOBACTERIA-ALPHA	erlit	3011
272568	*Gluconacetobacter diazotrophicus PAl* 5	PROTEOBACTERIA-ALPHA	gldia	7353
290633	*Gluconobacter oxydans* 621H	PROTEOBACTERIA-ALPHA	gloxy	2664
391165	*Granulibacter bethesdensis* CGDNIH1	PROTEOBACTERIA-ALPHA	grbet	2437
228405	*Hyphomonas neptunium ATCC* 15444	PROTEOBACTERIA-ALPHA	hynep	3505
290400	*Jannaschia* sp. CCS1	PROTEOBACTERIA-ALPHA	jasp	4283
156889	*Magnetococcus* sp. MC-1	PROTEOBACTERIA-ALPHA	masp	3716
394221	*Maricaulis maris* MCS10	PROTEOBACTERIA-ALPHA	mamar	3063
266835	*Mesorhizobium loti* MAFF303099	PROTEOBACTERIA-ALPHA	melot	7272
266779	*Mesorhizobium* sp. BNC1	PROTEOBACTERIA-ALPHA	mesp	4543
440085	*Methylobacterium chloromethanicum* CM4	PROTEOBACTERIA-ALPHA	mechl	5516
419610	*Methylobacterium extorquens* PA1	PROTEOBACTERIA-ALPHA	meext	4829
460265	*Methylobacterium nodulans ORS* 2060	PROTEOBACTERIA-ALPHA	menod	8308
441620	*Methylobacterium populi* BJ001	PROTEOBACTERIA-ALPHA	mepop	5365
426355	*Methylobacterium radiotolerans JCM* 2831	PROTEOBACTERIA-ALPHA	merad	6431
426117	*Methylobacterium* sp. 4-46	PROTEOBACTERIA-ALPHA	mesp	6692
395965	*Methylocella silvestris* BL2	PROTEOBACTERIA-ALPHA	mesil	3818
222891	*Neorickettsia sennetsu* str. Miyayama	PROTEOBACTERIA-ALPHA	nesen	932
323097	*Nitrobacter hamburgensis* X14	PROTEOBACTERIA-ALPHA	niham	4326
323098	*Nitrobacter winogradskyi* Nb-255	PROTEOBACTERIA-ALPHA	niwin	3122
279238	*Novosphingobium aromaticivorans DSM* 12444	PROTEOBACTERIA-ALPHA	noaro	3937
439375	*Ochrobactrum anthropi ATCC* 49188	PROTEOBACTERIA-ALPHA	ocant	4799
504832	*Oligotropha carboxidovorans* OM5	PROTEOBACTERIA-ALPHA	olcar	3722
357244	*Orientia tsutsugamushi* str. Boryong	PROTEOBACTERIA-ALPHA	ortsu	1182
334380	*Orientia tsutsugamushi* str. Ikeda	PROTEOBACTERIA-ALPHA	ortsu	1967
318586	*Paracoccus denitrificans* PD1222	PROTEOBACTERIA-ALPHA	paden	5077
402881	*Parvibaculum lavamentivorans* DS-1	PROTEOBACTERIA-ALPHA	palav	3636
450851	*Phenylobacterium zucineum* HLK1	PROTEOBACTERIA-ALPHA	phzuc	3854
347834	*Rhizobium etli CFN* 42	PROTEOBACTERIA-ALPHA	rhetl	5963
491916	*Rhizobium etli CIAT* 652	PROTEOBACTERIA-ALPHA	rhetl	6056
395492	*Rhizobium leguminosarum bv. trifolii* WSM2304	PROTEOBACTERIA-ALPHA	rhleg	6415
216596	*Rhizobium leguminosarum bv. viciae* 3841	PROTEOBACTERIA-ALPHA	rhleg	7143
272943	*Rhodobacter sphaeroides* 2.4.1	PROTEOBACTERIA-ALPHA	rhsph	4242
349102	*Rhodobacter sphaeroides ATCC* 17025	PROTEOBACTERIA-ALPHA	rhsph	4333
349101	*Rhodobacter sphaeroides ATCC* 17029	PROTEOBACTERIA-ALPHA	rhsph	4132
557760	*Rhodobacter sphaeroides* KD131	PROTEOBACTERIA-ALPHA	rhsph	4569
338969	*Rhodoferax ferrireducens* T118	PROTEOBACTERIA-ALPHA	rhfer	4418
316055	*Rhodopseudomonas palustris* BisA53	PROTEOBACTERIA-ALPHA	rhpal	4878
316056	*Rhodopseudomonas palustris* BisB18	PROTEOBACTERIA-ALPHA	rhpal	4886
316057	*Rhodopseudomonas palustris* BisB5	PROTEOBACTERIA-ALPHA	rhpal	4397
258594	*Rhodopseudomonas palustris* CGA009	PROTEOBACTERIA-ALPHA	rhpal	4820
316058	*Rhodopseudomonas palustris* HaA2	PROTEOBACTERIA-ALPHA	rhpal	4683
395960	*Rhodopseudomonas palustris* TIE-1	PROTEOBACTERIA-ALPHA	rhpal	5246
414684	*Rhodospirillum centenum SW (Rhodocista centenaria* SW)	PROTEOBACTERIA-ALPHA	rhcen	4002
269796	*Rhodospirillum rubrum ATCC* 11170	PROTEOBACTERIA-ALPHA	rhrub	3841
293614	*Rickettsia akari* str. Hartford	PROTEOBACTERIA-ALPHA	riaka	1259
391896	*Rickettsia bellii OSU* 85-389	PROTEOBACTERIA-ALPHA	ribel	1476
336407	*Rickettsia bellii* RML369-C	PROTEOBACTERIA-ALPHA	ribel	1429
293613	*Rickettsia canadensis* str. McKiel	PROTEOBACTERIA-ALPHA	rican	1093
272944	*Rickettsia conorii* str. Malish 7	PROTEOBACTERIA-ALPHA	ricon	1374
315456	*Rickettsia felis* URRWXCal2	PROTEOBACTERIA-ALPHA	rifel	1512
416276	*Rickettsia massiliae* MTU5	PROTEOBACTERIA-ALPHA	rimas	980
272947	*Rickettsia prowazekii* str. Madrid E	PROTEOBACTERIA-ALPHA	ripro	835
392021	*Rickettsia rickettsii* str. 'Sheila Smith'	PROTEOBACTERIA-ALPHA	riric	1345
452659	*Rickettsia rickettsii* str. Iowa	PROTEOBACTERIA-ALPHA	riric	1384
257363	*Rickettsia typhi* str. Wilmington	PROTEOBACTERIA-ALPHA	rityp	838
375451	*Roseobacter denitrificans OCh* 114	PROTEOBACTERIA-ALPHA	roden	4129
246200	*Silicibacter pomeroyi* DSS-3	PROTEOBACTERIA-ALPHA	sipom	4252
292414	*Silicibacter* sp. TM1040	PROTEOBACTERIA-ALPHA	sisp	3864

Taxonid	Organism name	Classification	Short name	Number of protein-coding genes
366394	*Sinorhizobium medicae* WSM419	PROTEOBACTERIA-ALPHA	simed	6213
266834	*Sinorhizobium meliloti* 1021	PROTEOBACTERIA-ALPHA	simel	6218
392499	*Sphingomonas wittichii* RW1	PROTEOBACTERIA-ALPHA	spwit	5345
317655	*Sphingopyxis alaskensis* RB2256	PROTEOBACTERIA-ALPHA	spala	3195
570417	*Wolbachia endosymbiont of Culex quinquefasciatus* Pel	PROTEOBACTERIA-ALPHA	woend	1275
163164	*Wolbachia endosymbiont of Drosophila* melanogaster	PROTEOBACTERIA-ALPHA	woend	1195
292805	*Wolbachia endosymbiont strain TRS of Brugia* malayi	PROTEOBACTERIA-ALPHA	woend	805
78245	*Xanthobacter autotrophicus* Py2	PROTEOBACTERIA-ALPHA	xaaut	5035
264203	*Zymomonas mobilis* subsp. mobilis ZM4	PROTEOBACTERIA-ALPHA	zymob	1998
397945	*Acidovorax avenae* subsp. citrulli AAC00-1	PROTEOBACTERIA-BETA	acave	4709
232721	*Acidovorax* sp. JS42	PROTEOBACTERIA-BETA	acsp	4155
76114	*Aromatoleum aromaticum* EbN1	PROTEOBACTERIA-BETA	araro	4590
62928	*Azoarcus* sp. BH72	PROTEOBACTERIA-BETA	azsp	3989
360910	*Bordetella avium* 197N	PROTEOBACTERIA-BETA	boavi	3381
257310	*Bordetella bronchiseptica* RB50	PROTEOBACTERIA-BETA	bobro	4994
257311	*Bordetella parapertussis* 12822	PROTEOBACTERIA-BETA	bopar	4185
257313	*Bordetella pertussis Tohama* I	PROTEOBACTERIA-BETA	boper	3436
340100	*Bordetella petrii DSM* 12804	PROTEOBACTERIA-BETA	bopet	5027
339670	*Burkholderia ambifaria* AMMD	PROTEOBACTERIA-BETA	buamb	6610
398577	*Burkholderia ambifaria* MC40-6	PROTEOBACTERIA-BETA	buamb	6697
331271	*Burkholderia cenocepacia AU* 1054	PROTEOBACTERIA-BETA	bucen	6477
331272	*Burkholderia cenocepacia* HI2424	PROTEOBACTERIA-BETA	bucen	6919
216591	*Burkholderia cenocepacia* J2315	PROTEOBACTERIA-BETA	bucen	7116
406425	*Burkholderia cenocepacia* MC0-3	PROTEOBACTERIA-BETA	bucen	7008
243160	*Burkholderia mallei ATCC* 23344	PROTEOBACTERIA-BETA	bumal	5024
412022	*Burkholderia mallei NCTC* 10229	PROTEOBACTERIA-BETA	bumal	5510
320389	*Burkholderia mallei NCTC* 10247	PROTEOBACTERIA-BETA	bumal	5852
320388	*Burkholderia mallei* SAVP1	PROTEOBACTERIA-BETA	bumal	5189
395019	*Burkholderia multivorans ATCC* 17616	PROTEOBACTERIA-BETA	bumul	12371
391038	*Burkholderia phymatum* STM815	PROTEOBACTERIA-BETA	buphy	7496
398527	*Burkholderia phytofirmans* PsJN	PROTEOBACTERIA-BETA	buphy	7241
357348	*Burkholderia pseudomallei* 1106a	PROTEOBACTERIA-BETA	bupse	7175
320372	*Burkholderia pseudomallei* 1710b	PROTEOBACTERIA-BETA	bupse	6347
320373	*Burkholderia pseudomallei* 668	PROTEOBACTERIA-BETA	bupse	7230
272560	*Burkholderia pseudomallei* K96243	PROTEOBACTERIA-BETA	bupse	5728
269483	*Burkholderia* sp. 383	PROTEOBACTERIA-BETA	busp	7717
271848	*Burkholderia thailandensis* E264	PROTEOBACTERIA-BETA	butha	5634
269482	*Burkholderia vietnamiensis* G4	PROTEOBACTERIA-BETA	buvie	7617
266265	*Burkholderia xenovorans* LB400	PROTEOBACTERIA-BETA	buxen	8702
243365	*Chromobacterium violaceum ATCC* 12472	PROTEOBACTERIA-BETA	chvio	4407
164546	*Cupriavidus* taiwanensis	PROTEOBACTERIA-BETA	cutai	5897
159087	*Dechloromonas aromatica* RCB	PROTEOBACTERIA-BETA	dearo	4171
398578	*Delftia acidovorans* SPH-1	PROTEOBACTERIA-BETA	deaci	6040
535289	*Diaphorobacter* sp. TPSY	PROTEOBACTERIA-BETA	disp	3479
204773	*Herminiimonas* arsenicoxydans	PROTEOBACTERIA-BETA	hears	3295
375286	*Janthinobacterium* sp. Marseille	PROTEOBACTERIA-BETA	jasp	3697
395495	*Leptothrix cholodnii* SP-6	PROTEOBACTERIA-BETA	lecho	4363
420662	*Methylibium petroleiphilum* PM1	PROTEOBACTERIA-BETA	mepet	4449
265072	*Methylobacillus flagellatus* KT	PROTEOBACTERIA-BETA	mefla	2753
242231	*Neisseria gonorrhoeae FA* 1090	PROTEOBACTERIA-BETA	negon	2002
521006	*Neisseria gonorrhoeae* NCCP11945	PROTEOBACTERIA-BETA	negon	2674
374833	*Neisseria meningitidis* 053442	PROTEOBACTERIA-BETA	nemen	2020
272831	*Neisseria meningitidis* FAM18	PROTEOBACTERIA-BETA	nemen	1917
122586	*Neisseria meningitidis* MC58	PROTEOBACTERIA-BETA	nemen	2063
122587	*Neisseria meningitidis* Z2491	PROTEOBACTERIA-BETA	nemen	1909
228410	*Nitrosomonas europaea ATCC* 19718	PROTEOBACTERIA-BETA	nieur	2461
335283	*Nitrosomonas eutropha* C91	PROTEOBACTERIA-BETA	nieut	2551
323848	*Nitrosospira multiformis ATCC* 25196	PROTEOBACTERIA-BETA	nimul	2805
365044	*Polaromonas naphthalenivorans* CJ2	PROTEOBACTERIA-BETA	ponap	4929
296591	*Polaromonas* sp. JS666	PROTEOBACTERIA-BETA	posp	5453
312153	*Polynucleobacter necessarius* subsp. asymbioticus QLW-P1DMWA-1	PROTEOBACTERIA-BETA	ponec	2077
452638	*Polynucleobacter necessarius* subsp. necessarius STIR1	PROTEOBACTERIA-BETA	ponec	1508

Taxonid	Organism name	Classification	Short name	Number of protein-coding genes
381666	*Ralstonia eutropha* H16	PROTEOBACTERIA-BETA	raeut	6626
264198	*Ralstonia eutropha* JMP134	PROTEOBACTERIA-BETA	raeut	6446
266264	*Ralstonia metallidurans* CH34	PROTEOBACTERIA-BETA	ramet	6319
402626	*Ralstonia pickettii* 12J	PROTEOBACTERIA-BETA	rapic	4952
267608	*Ralstonia solanacearum* GMI1000	PROTEOBACTERIA-BETA	rasol	5113
85643	*Thauera* sp. MZ1T	PROTEOBACTERIA-BETA	thsp	75
292415	*Thiobacillus denitrificans ATCC* 25259	PROTEOBACTERIA-BETA	thden	2827
391735	*Verminephrobacter eiseniae* EF01-2	PROTEOBACTERIA-BETA	veeis	4947
455488	*Anaeromyxobacter dehalogenans* 2CP-1	PROTEOBACTERIA-DELTA	andeh	4473
290397	*Anaeromyxobacter dehalogenans* 2CP-C	PROTEOBACTERIA-DELTA	andeh	4346
404589	*Anaeromyxobacter* sp. Fw109-5	PROTEOBACTERIA-DELTA	ansp	4466
447217	*Anaeromyxobacter* sp. K	PROTEOBACTERIA-DELTA	ansp	4457
264462	*Bdellovibrio bacteriovorus* HD100	PROTEOBACTERIA-DELTA	bdbac	3587
439235	*Desulfatibacillum alkenivorans* AK-01	PROTEOBACTERIA-DELTA	dealk	5252
96561	*Desulfococcus oleovorans* Hxd3	PROTEOBACTERIA-DELTA	deole	3265
177439	*Desulfotalea psychrophila* LSv54	PROTEOBACTERIA-DELTA	depsy	3234
525146	*Desulfovibrio desulfuricans* subsp. desulfuricans str. ATCC 27774	PROTEOBACTERIA-DELTA	dedes	2356
207559	*Desulfovibrio desulfuricans* subsp. desulfuricans str. G20	PROTEOBACTERIA-DELTA	dedes	3775
391774	*Desulfovibrio vulgaris* DP4	PROTEOBACTERIA-DELTA	devul	3091
883	*Desulfovibrio vulgaris* str. ’Miyazaki F’	PROTEOBACTERIA-DELTA	devul	3180
882	*Desulfovibrio vulgaris* str. Hildenborough	PROTEOBACTERIA-DELTA	devul	3531
404380	*Geobacter bemidjiensis* Bem	PROTEOBACTERIA-DELTA	gebem	4018
398767	*Geobacter lovleyi* SZ	PROTEOBACTERIA-DELTA	gelov	3685
269799	*Geobacter metallireducens* GS-15	PROTEOBACTERIA-DELTA	gemet	3532
316067	*Geobacter* sp. FRC-32	PROTEOBACTERIA-DELTA	gesp	3798
243231	*Geobacter sulfurreducens* PCA	PROTEOBACTERIA-DELTA	gesul	3445
351605	*Geobacter uraniireducens* Rf4	PROTEOBACTERIA-DELTA	geura	4357
363253	*Lawsonia intracellularis* PHE/MN1-00	PROTEOBACTERIA-DELTA	laint	1337
246197	*Myxococcus xanthus DK* 1622	PROTEOBACTERIA-DELTA	myxan	7331
338963	*Pelobacter carbinolicus DSM* 2380	PROTEOBACTERIA-DELTA	pecar	3352
338966	*Pelobacter propionicus DSM* 2379	PROTEOBACTERIA-DELTA	pepro	3804
448385	*Sorangium cellulosum ’So ce* 56’	PROTEOBACTERIA-DELTA	socel	9381
335543	*Syntrophobacter fumaroxidans* MPOB	PROTEOBACTERIA-DELTA	syfum	4064
56780	*Syntrophus aciditrophicus* SB	PROTEOBACTERIA-DELTA	syaci	3168
367737	*Arcobacter butzleri* RM4018	PROTEOBACTERIA-EPSILON	arbut	2259
360104	*Campylobacter concisus* 13826	PROTEOBACTERIA-EPSILON	cacon	1985
360105	*Campylobacter curvus* 525.92	PROTEOBACTERIA-EPSILON	cacur	1931
360106	*Campylobacter fetus* subsp. fetus 82-40	PROTEOBACTERIA-EPSILON	cafet	1719
360107	*Campylobacter hominis ATCC* BAA-381	PROTEOBACTERIA-EPSILON	cahom	1687
195099	*Campylobacter jejuni* RM1221	PROTEOBACTERIA-EPSILON	cajej	1838
360109	*Campylobacter jejuni* subsp. doylei 269.97	PROTEOBACTERIA-EPSILON	cajej	1731
354242	*Campylobacter jejuni* subsp. jejuni 81-176	PROTEOBACTERIA-EPSILON	cajej	1758
407148	*Campylobacter jejuni* subsp. jejuni 81116	PROTEOBACTERIA-EPSILON	cajej	1626
192222	*Campylobacter jejuni* subsp. jejuni NCTC 11168	PROTEOBACTERIA-EPSILON	cajej	1623
306263	*Campylobacter lari* RM2100	PROTEOBACTERIA-EPSILON	calar	1545
382638	*Helicobacter acinonychis* str. Sheeba	PROTEOBACTERIA-EPSILON	heaci	1618
235279	*Helicobacter hepaticus ATCC* 51449	PROTEOBACTERIA-EPSILON	hehep	1875
85962	*Helicobacter pylori* 26695	PROTEOBACTERIA-EPSILON	hepyl	1576
563041	*Helicobacter pylori* G27	PROTEOBACTERIA-EPSILON	hepyl	1504

Taxonid	Organism name	Classification	Short name	Number of protein-coding genes
357544	*Helicobacter pylori* HPAG1	PROTEOBACTERIA-EPSILON	hepyl	1544
85963	*Helicobacter pylori* J99	PROTEOBACTERIA-EPSILON	hepyl	1489
570508	*Helicobacter pylori* P12	PROTEOBACTERIA-EPSILON	hepyl	1578
512562	*Helicobacter pylori* Shi470	PROTEOBACTERIA-EPSILON	hepyl	1569
387092	*Nitratiruptor* sp. SB155-2	PROTEOBACTERIA-EPSILON	nisp	1843
326298	*Sulfurimonas denitrificans DSM* 1251	PROTEOBACTERIA-EPSILON	suden	2096
387093	*Sulfurovum* sp. NBC37-1	PROTEOBACTERIA-EPSILON	susp	2438
273121	*Wolinella succinogenes DSM* 1740	PROTEOBACTERIA-EPSILON	wosuc	2042
243159	*Acidithiobacillus ferrooxidans ATCC* 23270	PROTEOBACTERIA-GAMMA	acfer	3147
380394	*Acidithiobacillus ferrooxidans ATCC* 53993	PROTEOBACTERIA-GAMMA	acfer	2826
480119	*Acinetobacter baumannii* AB0057	PROTEOBACTERIA-GAMMA	acbau	3801
557600	*Acinetobacter baumannii* AB307-0294	PROTEOBACTERIA-GAMMA	acbau	3451
405416	*Acinetobacter baumannii* ACICU	PROTEOBACTERIA-GAMMA	acbau	3759
400667	*Acinetobacter baumannii ATCC* 17978	PROTEOBACTERIA-GAMMA	acbau	3367
509173	*Acinetobacter baumannii* AYE	PROTEOBACTERIA-GAMMA	acbau	3712
509170	*Acinetobacter baumannii* SDF	PROTEOBACTERIA-GAMMA	acbau	2975
62977	*Acinetobacter* sp. ADP1	PROTEOBACTERIA-GAMMA	acsp	3307
416269	*Actinobacillus pleuropneumoniae* L20	PROTEOBACTERIA-GAMMA	acple	2012
434271	*Actinobacillus pleuropneumoniae* serovar 3 str. JL03	PROTEOBACTERIA-GAMMA	acple	2036
537457	*Actinobacillus pleuropneumoniae* serovar 7 str. AP76	PROTEOBACTERIA-GAMMA	acple	2142
339671	*Actinobacillus succinogenes* 130Z	PROTEOBACTERIA-GAMMA	acsuc	2079
380703	*Aeromonas hydrophila* subsp. hydrophila ATCC 7966	PROTEOBACTERIA-GAMMA	aehyd	4122
382245	*Aeromonas salmonicida* subsp. salmonicida A449	PROTEOBACTERIA-GAMMA	aesal	4437
393595	*Alcanivorax borkumensis* SK2	PROTEOBACTERIA-GAMMA	albor	2755
316275	*Aliivibrio salmonicida* LFI1238	PROTEOBACTERIA-GAMMA	alsal	3911
187272	*Alkalilimnicola ehrlichei* MLHE-1	PROTEOBACTERIA-GAMMA	alehr	2865
314275	*Alteromonas macleodii 'Deep* ecotype'	PROTEOBACTERIA-GAMMA	almac	4072
374463	*Baumannia cicadellinicola* str. Hc (Homalodisca coagulata)	PROTEOBACTERIA-GAMMA	bacic	595
261318	*Buchnera aphidicola (Cinara* cedri)	PROTEOBACTERIA-GAMMA	buaph	5
563178	*Buchnera aphidicola* str. 5A (Acyrthosiphon pisum)	PROTEOBACTERIA-GAMMA	buaph	555
107806	*Buchnera aphidicola* str. APS (Acyrthosiphon pisum)	PROTEOBACTERIA-GAMMA	buaph	574
224915	*Buchnera aphidicola* str. Bp (Baizongia pistaciae)	PROTEOBACTERIA-GAMMA	buaph	507
372461	*Buchnera aphidicola* str. Cc (Cinara cedri)	PROTEOBACTERIA-GAMMA	buaph	357

Taxonid	Organism name	Classification	Short name	Number of protein-coding genes
198804	*Buchnera aphidicola* str. Sg (Schizaphis graminum)	PROTEOBACTERIA-GAMMA	buaph	546
561501	*Buchnera aphidicola* str. Tuc7 (Acyrthosiphon pisum)	PROTEOBACTERIA-GAMMA	buaph	553
203907	*Candidatus Blochmannia* floridanus	PROTEOBACTERIA-GAMMA	cablo	583
291272	*Candidatus Blochmannia pennsylvanicus* str. BPEN	PROTEOBACTERIA-GAMMA	cablo	610
387662	*Candidatus Carsonella ruddii* PV	PROTEOBACTERIA-GAMMA	cacar	182
413404	*Candidatus Ruthia magnifica* str. Cm (Calyptogena magnifica)	PROTEOBACTERIA-GAMMA	carut	976
412965	*Candidatus Vesicomyosocius okutanii HA (Candidatus Vesicomyosocius okutanii* str. HA)	PROTEOBACTERIA-GAMMA	caves	937
498211	*Cellvibrio japonicus* Ueda107	PROTEOBACTERIA-GAMMA	cejap	3754
290398	*Chromohalobacter salexigens DSM* 3043	PROTEOBACTERIA-GAMMA	chsal	3298
290338	*Citrobacter koseri ATCC* BAA-895	PROTEOBACTERIA-GAMMA	cikos	5008
167879	*Colwellia psychrerythraea* 34H	PROTEOBACTERIA-GAMMA	copsy	4910
434923	*Coxiella burnetii* CbuG_Q212	PROTEOBACTERIA-GAMMA	cobur	1871
434924	*Coxiella burnetii* CbuK_Q154	PROTEOBACTERIA-GAMMA	cobur	1947
434922	*Coxiella burnetii Dugway* 5J108-111	PROTEOBACTERIA-GAMMA	cobur	2045
360115	*Coxiella burnetii RSA* 331	PROTEOBACTERIA-GAMMA	cobur	1975
227377	*Coxiella burnetii RSA* 493	PROTEOBACTERIA-GAMMA	cobur	1848
246195	*Dichelobacter nodosus* VCS1703A	PROTEOBACTERIA-GAMMA	dinod	1280
290339	*Enterobacter sakazakii ATCC* BAA-894	PROTEOBACTERIA-GAMMA	ensak	4420
399742	*Enterobacter* sp. 638	PROTEOBACTERIA-GAMMA	ensp	4240
465817	*Erwinia tasmaniensis* Et1/99	PROTEOBACTERIA-GAMMA	ertas	3622
362663	*Escherichia coli* 536	PROTEOBACTERIA-GAMMA	escol	4620
585055	*Escherichia coli* 55989	PROTEOBACTERIA-GAMMA	escol	4763
405955	*Escherichia coli APEC* O1	PROTEOBACTERIA-GAMMA	escol	4851
481805	*Escherichia coli ATCC* 8739	PROTEOBACTERIA-GAMMA	escol	4200
199310	*Escherichia coli* CFT073	PROTEOBACTERIA-GAMMA	escol	5339
331111	*Escherichia coli* E24377A	PROTEOBACTERIA-GAMMA	escol	4991
585397	*Escherichia coli* ED1a	PROTEOBACTERIA-GAMMA	escol	4915
331112	*Escherichia coli* HS	PROTEOBACTERIA-GAMMA	escol	4378
585034	*Escherichia coli* IAI1	PROTEOBACTERIA-GAMMA	escol	4353
585057	*Escherichia coli* IAI39	PROTEOBACTERIA-GAMMA	escol	4732
591946	*Escherichia coli* LF82	PROTEOBACTERIA-GAMMA	escol	4312
574521	*Escherichia coli O127:H6* str. E2348/69	PROTEOBACTERIA-GAMMA	escol	4653
155864	*Escherichia coli O157:H7* EDL933	PROTEOBACTERIA-GAMMA	escol	5411

Taxonid	Organism name	Classification	Short name	Number of protein-coding genes
444450	*Escherichia coli O157:H7* str. EC4115	PROTEOBACTERIA-GAMMA	escol	5477
386585	*Escherichia coli O157:H7* str. Sakai	PROTEOBACTERIA-GAMMA	escol	5318
585035	*Escherichia coli* S88	PROTEOBACTERIA-GAMMA	escol	4696
409438	*Escherichia coli* SE11	PROTEOBACTERIA-GAMMA	escol	5002
439855	*Escherichia coli* SMS-3-5	PROTEOBACTERIA-GAMMA	escol	4913
316385	*Escherichia coli* str. K-12 substr. DH10B	PROTEOBACTERIA-GAMMA	escol	4126
511145	*Escherichia coli* str. K-12 substr. MG1655	PROTEOBACTERIA-GAMMA	escol	4131
316407	*Escherichia coli* str. K-12 substr. W3110	PROTEOBACTERIA-GAMMA	escol	4226
585056	*Escherichia coli* UMN026	PROTEOBACTERIA-GAMMA	escol	4968
364106	*Escherichia coli* UTI89	PROTEOBACTERIA-GAMMA	escol	5166
401614	*Francisella novicida* U112	PROTEOBACTERIA-GAMMA	frnov	1719
484022	*Francisella philomiragia* subsp. philomiragia ATCC 25017	PROTEOBACTERIA-GAMMA	frphi	1915
119857	*Francisella tularensis* subsp. holarctica	PROTEOBACTERIA-GAMMA	frtul	1754
458234	*Francisella tularensis* subsp. holarctica FTNF002-00	PROTEOBACTERIA-GAMMA	frtul	1580
393011	*Francisella tularensis* subsp. holarctica OSU18	PROTEOBACTERIA-GAMMA	frtul	1555
441952	*Francisella tularensis* subsp. mediasiatica FSC147	PROTEOBACTERIA-GAMMA	frtul	1406
393115	*Francisella tularensis* subsp. tularensis FSC198	PROTEOBACTERIA-GAMMA	frtul	1605
177416	*Francisella tularensis* subsp. tularensis SCHU S4	PROTEOBACTERIA-GAMMA	frtul	1603
418136	*Francisella tularensis* subsp. tularensis WY96-3418	PROTEOBACTERIA-GAMMA	frtul	1634
233412	*Haemophilus ducreyi* 35000HP	PROTEOBACTERIA-GAMMA	haduc	1717
281310	*Haemophilus influenzae* 86-028NP	PROTEOBACTERIA-GAMMA	hainf	1792
374930	*Haemophilus influenzae* PittEE	PROTEOBACTERIA-GAMMA	hainf	1619
374931	*Haemophilus influenzae* PittGG	PROTEOBACTERIA-GAMMA	hainf	1667
71421	*Haemophilus influenzae Rd* KW20	PROTEOBACTERIA-GAMMA	hainf	1657
557723	*Haemophilus parasuis* SH0165	PROTEOBACTERIA-GAMMA	hapar	2021
205914	*Haemophilus somnus* 129PT	PROTEOBACTERIA-GAMMA	hasom	1798
228400	*Haemophilus somnus* 2336	PROTEOBACTERIA-GAMMA	hasom	1980
349521	*Hahella chejuensis KCTC* 2396	PROTEOBACTERIA-GAMMA	hache	6778
349124	*Halorhodospira halophila* SL1	PROTEOBACTERIA-GAMMA	hahal	2407
283942	*Idiomarina loihiensis* L2TR	PROTEOBACTERIA-GAMMA	idloi	2628
507522	*Klebsiella pneumoniae* 342	PROTEOBACTERIA-GAMMA	klpne	5768
272620	*Klebsiella pneumoniae* subsp. pneumoniae MGH 78578	PROTEOBACTERIA-GAMMA	klpne	5185
400673	*Legionella pneumophila* str. Corby	PROTEOBACTERIA-GAMMA	lepne	3206

Taxonid	Organism name	Classification	Short name	Number of protein-coding genes
297245	*Legionella pneumophila* str. Lens	PROTEOBACTERIA-GAMMA	lepne	2934
297246	*Legionella pneumophila* str. Paris	PROTEOBACTERIA-GAMMA	lepne	3166
272624	*Legionella pneumophila* subsp. pneumophila str. Philadelphia 1	PROTEOBACTERIA-GAMMA	lepne	2942
221988	*Mannheimia succiniciproducens* MBEL55E	PROTEOBACTERIA-GAMMA	masuc	2369
351348	*Marinobacter aquaeolei* VT8	PROTEOBACTERIA-GAMMA	maaqu	4272
400668	*Marinomonas* sp. MWYL1	PROTEOBACTERIA-GAMMA	masp	4439
243233	*Methylococcus capsulatus* str. Bath	PROTEOBACTERIA-GAMMA	mecap	2956
323261	*Nitrosococcus oceani ATCC* 19707	PROTEOBACTERIA-GAMMA	nioce	3017
272843	*Pasteurella multocida* subsp. multocida str. Pm70	PROTEOBACTERIA-GAMMA	pamul	2015
218491	*Pectobacterium atrosepticum* SCRI1043	PROTEOBACTERIA-GAMMA	peatr	4472
298386	*Photobacterium profundum* SS9	PROTEOBACTERIA-GAMMA	phpro	5489
243265	*Photorhabdus luminescens* subsp. laumondii TTO1	PROTEOBACTERIA-GAMMA	phlum	4683
529507	*Proteus mirabilis* HI4320	PROTEOBACTERIA-GAMMA	prmir	3662
342610	*Pseudoalteromonas atlantica* T6c	PROTEOBACTERIA-GAMMA	psatl	4281
326442	*Pseudoalteromonas haloplanktis* TAC125	PROTEOBACTERIA-GAMMA	pshal	3485
557722	*Pseudomonas aeruginosa* LESB58	PROTEOBACTERIA-GAMMA	psaer	5925
381754	*Pseudomonas aeruginosa* PA7	PROTEOBACTERIA-GAMMA	psaer	6286
208964	*Pseudomonas aeruginosa* PAO1	PROTEOBACTERIA-GAMMA	psaer	5566
208963	*Pseudomonas aeruginosa* UCBPP-PA14	PROTEOBACTERIA-GAMMA	psaer	5892
384676	*Pseudomonas entomophila* L48	PROTEOBACTERIA-GAMMA	psent	5134
220664	*Pseudomonas fluorescens* Pf-5	PROTEOBACTERIA-GAMMA	psflu	6138
205922	*Pseudomonas fluorescens* Pf0-1	PROTEOBACTERIA-GAMMA	psflu	5736
399739	*Pseudomonas mendocina* ymp	PROTEOBACTERIA-GAMMA	psmen	4594
351746	*Pseudomonas putida* F1	PROTEOBACTERIA-GAMMA	psput	5252
76869	*Pseudomonas putida* GB-1	PROTEOBACTERIA-GAMMA	psput	5409
160488	*Pseudomonas putida* KT2440	PROTEOBACTERIA-GAMMA	psput	5350
390235	*Pseudomonas putida* W619	PROTEOBACTERIA-GAMMA	psput	5182
379731	*Pseudomonas stutzeri* A1501	PROTEOBACTERIA-GAMMA	psstu	4128
264730	*Pseudomonas syringae pv. phaseolicola* 1448A	PROTEOBACTERIA-GAMMA	pssyr	5172
205918	*Pseudomonas syringae pv. syringae* B728a	PROTEOBACTERIA-GAMMA	pssyr	5089
223283	*Pseudomonas syringae pv. tomato* str. DC3000	PROTEOBACTERIA-GAMMA	pssyr	5614
259536	*Psychrobacter arcticus* 273-4	PROTEOBACTERIA-GAMMA	psarc	2120
335284	*Psychrobacter cryohalolentis* K5	PROTEOBACTERIA-GAMMA	pscry	2511

Taxonid	Organism name	Classification	Short name	Number of protein-coding genes
349106	*Psychrobacter* sp. PRwf-1	PROTEOBACTERIA-GAMMA	pssp	2385
357804	*Psychromonas ingrahamii* 37	PROTEOBACTERIA-GAMMA	psing	3545
203122	*Saccharophagus degradans* 2-40	PROTEOBACTERIA-GAMMA	sadeg	4007
41514	*Salmonella enterica* subsp. arizonae serovar 62:z4,z23:–	PROTEOBACTERIA-GAMMA	saent	4498
454166	*Salmonella enterica* subsp. enterica serovar Agona str. SL483	PROTEOBACTERIA-GAMMA	saent	4614
321314	*Salmonella enterica* subsp. enterica serovar Choleraesuis str. SC-B67	PROTEOBACTERIA-GAMMA	saent	4634
439851	*Salmonella enterica* subsp. enterica serovar Dublin str. CT_02021853	PROTEOBACTERIA-GAMMA	saent	4617
550537	*Salmonella enterica* subsp. enterica serovar Enteritidis str. P125109	PROTEOBACTERIA-GAMMA	saent	4206
550538	*Salmonella enterica* subsp. enterica serovar Gallinarum str. 287/91	PROTEOBACTERIA-GAMMA	saent	3965
454169	*Salmonella enterica* subsp. enterica serovar Heidelberg str. SL476	PROTEOBACTERIA-GAMMA	saent	4779
423368	*Salmonella enterica* subsp. enterica serovar Newport str. SL254	PROTEOBACTERIA-GAMMA	saent	4805
554290	*Salmonella enterica* subsp. enterica serovar Paratyphi A str. AKU_12601	PROTEOBACTERIA-GAMMA	saent	4078
295319	*Salmonella enterica* subsp. enterica serovar Paratyphi A str. ATCC 9150	PROTEOBACTERIA-GAMMA	saent	4093
272994	*Salmonella enterica* subsp. enterica serovar Paratyphi B str. SPB7	PROTEOBACTERIA-GAMMA	saent	5592
439843	*Salmonella enterica* subsp. enterica serovar Schwarzengrund str. CVM19633	PROTEOBACTERIA-GAMMA	saent	4627
220341	*Salmonella enterica* subsp. enterica serovar Typhi str. CT18	PROTEOBACTERIA-GAMMA	saent	4758
209261	*Salmonella enterica* subsp. enterica serovar Typhi str. Ty2	PROTEOBACTERIA-GAMMA	saent	4318
99287	*Salmonella enterica* subsp. enterica serovar Typhimurium str. LT2	PROTEOBACTERIA-GAMMA	saent	102
128975	*Salmonella enterica* subsp. enterica serovar Typhimurium str. LT2	PROTEOBACTERIA-GAMMA	saent	4423
399741	*Serratia proteamaculans* 568	PROTEOBACTERIA-GAMMA	sepro	4942
326297	*Shewanella amazonensis* SB2B	PROTEOBACTERIA-GAMMA	shama	3645
325240	*Shewanella baltica* OS155	PROTEOBACTERIA-GAMMA	shbal	4489
402882	*Shewanella baltica* OS185	PROTEOBACTERIA-GAMMA	shbal	4394
399599	*Shewanella baltica* OS195	PROTEOBACTERIA-GAMMA	shbal	4688
407976	*Shewanella baltica* OS223	PROTEOBACTERIA-GAMMA	shbal	4441
318161	*Shewanella denitrificans* OS217	PROTEOBACTERIA-GAMMA	shden	3754
318167	*Shewanella frigidimarina NCIMB* 400	PROTEOBACTERIA-GAMMA	shfri	4029
458817	*Shewanella halifaxensis* HAW-EB4	PROTEOBACTERIA-GAMMA	shhal	4278
323850	*Shewanella loihica* PV-4	PROTEOBACTERIA-GAMMA	shloi	3859
211586	*Shewanella oneidensis* MR-1	PROTEOBACTERIA-GAMMA	shone	4467
398579	*Shewanella pealeana ATCC* 700345	PROTEOBACTERIA-GAMMA	shpea	4241
409026	*Shewanella piezotolerans* WP3	PROTEOBACTERIA-GAMMA	shpie	4933
319224	*Shewanella putrefaciens* CN-32	PROTEOBACTERIA-GAMMA	shput	3972

Taxonid	Organism name	Classification	Short name	Number of protein-coding genes
425104	*Shewanella sediminis* HAW-EB3	PROTEOBACTERIA-GAMMA	shsed	4497
94122	*Shewanella* sp. ANA-3	PROTEOBACTERIA-GAMMA	shsp	4360
60480	*Shewanella* sp. MR-4	PROTEOBACTERIA-GAMMA	shsp	3924
60481	*Shewanella* sp. MR-7	PROTEOBACTERIA-GAMMA	shsp	4014
351745	*Shewanella* sp. W3-18-1	PROTEOBACTERIA-GAMMA	shsp	4044
392500	*Shewanella woodyi ATCC* 51908	PROTEOBACTERIA-GAMMA	shwoo	4880
344609	*Shigella boydii CDC* 3083-94	PROTEOBACTERIA-GAMMA	shboy	4557
300268	*Shigella boydii* Sb227	PROTEOBACTERIA-GAMMA	shboy	4282
300267	*Shigella dysenteriae* Sd197	PROTEOBACTERIA-GAMMA	shdys	4502
198215	*Shigella flexneri 2a* str. 2457T	PROTEOBACTERIA-GAMMA	shfle	4061
198214	*Shigella flexneri 2a* str. 301	PROTEOBACTERIA-GAMMA	shfle	4440
373384	*Shigella flexneri 5* str. 8401	PROTEOBACTERIA-GAMMA	shfle	4115
300269	*Shigella sonnei* Ss046	PROTEOBACTERIA-GAMMA	shson	4471
343509	*Sodalis glossinidius* str. 'morsitans'	PROTEOBACTERIA-GAMMA	soglo	2516
522373	*Stenotrophomonas maltophilia* K279a	PROTEOBACTERIA-GAMMA	stmal	4386
391008	*Stenotrophomonas maltophilia* R551-3	PROTEOBACTERIA-GAMMA	stmal	4039
396588	*Thioalkalivibrio* sp. HL-EbGR7	PROTEOBACTERIA-GAMMA	thsp	3283
317025	*Thiomicrospira crunogena* XCL-2	PROTEOBACTERIA-GAMMA	thcru	2196
243277	*Vibrio cholerae O1 biovar eltor* str. N16961	PROTEOBACTERIA-GAMMA	vicho	3835
345073	*Vibrio cholerae* O395	PROTEOBACTERIA-GAMMA	vicho	3875
312309	*Vibrio fischeri* ES114	PROTEOBACTERIA-GAMMA	vifis	3818
388396	*Vibrio fischeri* MJ11	PROTEOBACTERIA-GAMMA	vifis	4039
338187	*Vibrio harveyi ATCC* BAA-1116	PROTEOBACTERIA-GAMMA	vihar	6040
223926	*Vibrio parahaemolyticus RIMD* 2210633	PROTEOBACTERIA-GAMMA	vipar	4832
575788	*Vibrio splendidus* LGP32	PROTEOBACTERIA-GAMMA	vispl	4431
216895	*Vibrio vulnificus* CMCP6	PROTEOBACTERIA-GAMMA	vivul	4472
196600	*Vibrio vulnificus* YJ016	PROTEOBACTERIA-GAMMA	vivul	5024
36870	*Wigglesworthia glossinidia endosymbiont of Glossina* brevipalpis	PROTEOBACTERIA-GAMMA	wiglo	617
190486	*Xanthomonas axonopodis pv. citri* str. 306	PROTEOBACTERIA-GAMMA	xaaxo	4427
314565	*Xanthomonas campestris pv. campestris* str. 8004	PROTEOBACTERIA-GAMMA	xacam	4273
190485	*Xanthomonas campestris pv. campestris* str. ATCC 33913	PROTEOBACTERIA-GAMMA	xacam	4181
509169	*Xanthomonas campestris pv. campestris* str. B100	PROTEOBACTERIA-GAMMA	xacam	4467
316273	*Xanthomonas campestris pv. vesicatoria* str. 85-10	PROTEOBACTERIA-GAMMA	xacam	4726

Taxonid	Organism name	Classification	Short name	Number of protein-coding genes
291331	*Xanthomonas oryzae pv. oryzae* KACC10331	PROTEOBACTERIA-GAMMA	xaory	4062
342109	*Xanthomonas oryzae pv. oryzae MAFF* 311018	PROTEOBACTERIA-GAMMA	xaory	4372
360094	*Xanthomonas oryzae pv. oryzae* PXO99A	PROTEOBACTERIA-GAMMA	xaory	4988
160492	*Xylella fastidiosa* 9a5c	PROTEOBACTERIA-GAMMA	xyfas	2832
405440	*Xylella fastidiosa* M12	PROTEOBACTERIA-GAMMA	xyfas	2104
405441	*Xylella fastidiosa* M23	PROTEOBACTERIA-GAMMA	xyfas	2201
183190	*Xylella fastidiosa* Temecula1	PROTEOBACTERIA-GAMMA	xyfas	2036
393305	*Yersinia enterocolitica* subsp. enterocolitica 8081	PROTEOBACTERIA-GAMMA	yeent	4051
349746	*Yersinia pestis* Angola	PROTEOBACTERIA-GAMMA	yepes	4040
360102	*Yersinia pestis* Antiqua	PROTEOBACTERIA-GAMMA	yepes	4364
229193	*Yersinia pestis biovar Microtus* str. 91001	PROTEOBACTERIA-GAMMA	yepes	4138
214092	*Yersinia pestis* CO92	PROTEOBACTERIA-GAMMA	yepes	4066
187410	*Yersinia pestis* KIM	PROTEOBACTERIA-GAMMA	yepes	4202
377628	*Yersinia pestis* Nepal516	PROTEOBACTERIA-GAMMA	yepes	4094
386656	*Yersinia pestis Pestoides* F	PROTEOBACTERIA-GAMMA	yepes	4069
349747	*Yersinia pseudotuberculosis IP* 31758	PROTEOBACTERIA-GAMMA	yepse	4324
273123	*Yersinia pseudotuberculosis IP* 32953	PROTEOBACTERIA-GAMMA	yepse	4038
502801	*Yersinia pseudotuberculosis* PB1/+	PROTEOBACTERIA-GAMMA	yepse	4237
502800	*Yersinia pseudotuberculosis* YPIII	PROTEOBACTERIA-GAMMA	yepse	4192
333668	*apicoplast Theileria parva strain* Muguga	PROTISTS-APICOMPLEXA	apthe	2223
353152	*Cryptosporidium parvum Iowa* II	PROTISTS-APICOMPLEXA	crpar	3805
36329	*Plasmodium falciparum* 3D7	PROTISTS-APICOMPLEXA	plfal	5262
126793	*Plasmodium vivax* SaI-1	PROTISTS-APICOMPLEXA	plviv	5050
184922	*Giardia lamblia ATCC 50803 (Giardia intestinalis ATCC* 50803)	PROTISTS-DIPLOMONADIDA	gilam	6503
370354	*Entamoeba dispar* SAW760	PROTISTS-ENTAMOEBIDAE	endis	8812
294381	*Entamoeba histolytica* HM-1:IMSS	PROTISTS-ENTAMOEBIDAE	enhis	8163
420245	*Leishmania braziliensis* MHOM/BR/75/M2904	PROTISTS-EUGLENOZOA	lebra	7896
435258	*Leishmania infantum* JPCM5	PROTISTS-EUGLENOZOA	leinf	7993
347515	*Leishmania major strain* Friedlin	PROTISTS-EUGLENOZOA	lemaj	8265
185431	*Trypanosoma brucei* TREU927	PROTISTS-EUGLENOZOA	trbru	9279
352472	*Dictyostelium discoideum* AX4	PROTISTS-MYCETOZOA	didis	13331
412133	*Trichomonas vaginalis* G3	PROTISTS-PARABASALIDEA	trvag	59518
390236	*Borrelia afzelii* PKo	SPIROCHAETES	boafz	1214
224326	*Borrelia burgdorferi* B31	SPIROCHAETES	bobur	1640
445985	*Borrelia burgdorferi* ZS7	SPIROCHAETES	bobur	1239
412419	*Borrelia duttonii* Ly	SPIROCHAETES	bodut	1305
290434	*Borrelia garinii* PBi	SPIROCHAETES	bogar	1270
314723	*Borrelia hermsii* DAH	SPIROCHAETES	boher	819
412418	*Borrelia recurrentis* A1	SPIROCHAETES	borec	990
314724	*Borrelia turicatae* 91E135	SPIROCHAETES	botur	818
355278	*Leptospira biflexa* serovar Patoc strain ’Patoc 1 (Ames)’	SPIROCHAETES	lebif	3600
456481	*Leptospira biflexa* serovar Patoc strain ’Patoc 1 (Paris)’	SPIROCHAETES	lebif	3726

Taxonid	Organism name	Classification	Short name	Number of protein-coding genes
355277	*Leptospira borgpetersenii* serovar Hardjo-bovis JB197	SPIROCHAETES	lebor	2880
355276	*Leptospira borgpetersenii* serovar Hardjo-bovis L550	SPIROCHAETES	lebor	2945
267671	*Leptospira interrogans* serovar Copenhageni str. Fiocruz L1-130	SPIROCHAETES	leint	3658
189518	*Leptospira interrogans* serovar Lai str. 56601	SPIROCHAETES	leint	4724
243275	*Treponema denticola ATCC* 35405	SPIROCHAETES	trden	2767
455434	*Treponema pallidum* subsp. pallidum SS14	SPIROCHAETES	trpal	1028
243276	*Treponema pallidum* subsp. pallidum str. Nichols	SPIROCHAETES	trpal	1036
347257	*Mycoplasma agalactiae* PG2	TENERICUTES	myaga	742
243272	*Mycoplasma arthritidis* 158L3-1	TENERICUTES	myart	631
340047	*Mycoplasma capricolum* subsp. capricolum ATCC 27343	TENERICUTES	mycap	812
233150	*Mycoplasma gallisepticum* R	TENERICUTES	mygal	726
243273	*Mycoplasma genitalium* G37	TENERICUTES	mygen	476
295358	*Mycoplasma hyopneumoniae* 232	TENERICUTES	myhyo	691
262722	*Mycoplasma hyopneumoniae* 7448	TENERICUTES	myhyo	657
262719	*Mycoplasma hyopneumoniae* J	TENERICUTES	myhyo	657
267748	*Mycoplasma mobile* 163K	TENERICUTES	mymob	633
272632	*Mycoplasma mycoides* subsp. mycoides SC str. PG1	TENERICUTES	mymyc	1016
272633	*Mycoplasma penetrans* HF-2	TENERICUTES	mypen	1037
272634	*Mycoplasma pneumoniae* M129	TENERICUTES	mypne	689
272635	*Mycoplasma pulmonis UAB* CTIP	TENERICUTES	mypul	782
262723	*Mycoplasma synoviae* 53	TENERICUTES	mysyn	659
319795	*Deinococcus geothermalis DSM* 11300	THERMI	degeo	3054
243230	*Deinococcus radiodurans* R1	THERMI	derad	3167
262724	*Thermus thermophilus* HB27	THERMI	ththe	2210
300852	*Thermus thermophilus* HB8	THERMI	ththe	2238
381764	*Fervidobacterium nodosum* Rt17-B1	THERMOTOGAE	fenod	1750
403833	*Petrotoga mobilis* SJ95	THERMOTOGAE	pemob	1898
484019	*Thermosipho africanus* TCF52B	THERMOTOGAE	thafr	1911
391009	*Thermosipho melanesiensis* BI429	THERMOTOGAE	thmel	1879
416591	*Thermotoga lettingae* TMO	THERMOTOGAE	thlet	2040
243274	*Thermotoga maritima* MSB8	THERMOTOGAE	thmar	1858
309803	*Thermotoga neapolitana DSM* 4359	THERMOTOGAE	thnea	1937
390874	*Thermotoga petrophila* RKU-1	THERMOTOGAE	thpet	1785
126740	*Thermotoga* sp. RQ2	THERMOTOGAE	thsp	1819
349741	*Akkermansia muciniphila ATCC* BAA-835	VERRUCOMICROBIA	akmuc	2138
481448	*Methylacidiphilum infernorum* V4	VERRUCOMICROBIA	meinf	2472
452637	*Opitutus terrae* PB90-1	VERRUCOMICROBIA	opter	4612

Appendix B

Manual microbe-habitat annotation results from literature

The microorganism-habitat annotation results from manual exploration of 20 publications. Numerous host- and non host-associated microorganisms were investigated to determine information about their corresponding isolation sources or habitats from the literature. Terms used to describe relations between organisms and habitats were found to be highly variable. Terms referring to habitats were also found to vary substantially. Interestingly, highly specific terms such as ‘colonise’, ‘attach to’, and ‘survive in’ were used with human-related microorganisms, whereas much more generic terms including ‘isolated from’, ‘is’ (soil bacteria) and ‘grow in’ were commonly used to describe other microorganisms, particularly for non-host associated microbes.

Directly habitat-related sentence	**organisms**	**relationship**	**habitats**	**PMID**
Escherichia coli colonizes the lower gut of animals, and survives when released to the natural environment.	*Escherichia coli*	colonizes	lower gut of animals	9278503
		survives in	natural environment	
chlamydiae have an extremely broad host range (comprising protozoa, arthropods, and marsupial and placental mammals) and a ubiquitous, worldwide distribution in nature	chlamydiae	have host	protozoa, arthropods, marsupial mammals, placental mammals	15632447
large number of rRNA sequences detected in various clinical and environmental samples (including bronchoalveolar lavage, nose, throat and ocular swabs from humans and animals, fresh water, soil, and activated-sludge samples) represent as yet unknown chlamydiae, indicating that chlamydial diversity is still dramatically underestimated	chlamydiae	detected in	bronchoalveolar lavage, nose, throat, ocular swabs from human, ocular swabs from animals, fresh water, soil, activated-sludge samples	15632447
its successful colonization of the gastric environment	*Helicobacter pylori*	colonizes	gastric environment	9252185
The presence of the bacterium in the gastric mucosa is associated with chronic active gastritis. . .		presented in	gastric mucosa	9252185

Directly habitat-related sentence	**organisms**	**relationship**	**habitats**	**PMID**
Haemophilus influenzae natural host is human.	*Haemophilus influenzae*	have host	human	7542800
Non-typeable strains also exist and are distinguished by their lack of detectable capsular polysaccharide. They are commensal residents of the upper respiratory mucosa of children and adults and cause otitis media and respiratory tract infections.	*Haemophilus influenzae* (NTHi)	residents of	upper respiratory mucosa of children and adults	7542800
H. influenzae enters the body through respiration and establishes either an asymptomatic colonization or a frank infectious process within the host respiratory mucosa	*Haemophilus influenzae*	colonizes	host respiratory mucosa	15908377
Trichomonas vaginalis, The extracellular parasite resides in the urogenital tract of both sexes and can cause vaginitis in women and urethritis and prostatitis in men	*Trichomonas vaginalis*	resides in	urogenital tract of both sexes	17218520
Successful colonization of the host mucosa by *T. vaginalis* is thought to depend on multiple mechanisms including	*Trichomonas vaginalis*	colonizes	host mucosa	17962075
The protist *Trichomonas vaginalis* is one of the most common human sexually transmitted pathogens that colonize the urogenital mucosa.		colonizes	urogenital mucosa	17962075
Flagellated giardial trophozoites attach to epithelial cells of the small intestine, where they can cause disease without triggering a pronounced inflammatory response	*Giardia lamblia*	attach to	epithelial cells of the small intestine	17901334
Meningococcal infection starts with colonization of the nasopharyngeal and tonsillar mucosa.	*Neisseria meningitidis* (serogroup B)	colonizes	nasopharyngeal mucosa, tonsillar mucosa	12486052
Following adherence, meningococci initiate endocytosis, cross the epithelial barrier, and gain access to the bloodstream	*Neisseria meningitidis*	adheres	epithelial barrier	12486052
		access to	blood stream	12486052
Meningococci are characterized by a marked tropism towards the central nervous system	*Neisseria meningitidis*	tropism towards	central nervous system	12486052
increasing number of scientific reports describing adhesion of Lactobacillus to components of the human intestinal mucosa	Lactobacillus	adheres	human intestinal mucosa	17888009
Lactobacilli colonize the gastrointestinal and urinary tract of humans	Lactobacillus	colonizes	urinary tract of humans	17888009
Lactobacilli commonly found in the mouth, GI tract and female GUT	Lactobacillus	found in	mouth	17888009
			female genitourinary tract	17888009
Clostridium difficile, which can form longer-term relationships with its host. Overall, the genome indicates that *C. botulinum* is adapted to a saprophytic lifestyle both in soil and aquatic environments	*Clostridium botulinum*	survives in	soil and aquatic environments	17519437
Acinetobacter spp. can be obtained from water, soil and living organisms	*Acinetobacter spp.*	obtained from	water, soil, living organisms	15514110
Acinetobacter sp. strain ADP1 is a nutritionally versatile soil bacterium	*Acinetobacter sp.* strain ADP1	is	soil bacterium	15514110
Candidatus Phytoplasma can efficiently invade cells of insects and plants	*Candidatus Phytoplasma*	invade	cells of insects and plants	16672622
Phytoplasmas are generally associated with arthropods and plants	Phytoplasmas	associated with	arthropods and plants	16672622
Cyanothece sp. 113 was isolated from the sea in China	*Cyanothece sp.* 113	isolated from	sea	16782333
Arthrobacter spp. are very widely distributed in the environment (e.g., soil)	*Arthrobacter spp.*	distributed in	environment (e.g., soil)	8880479
TABLE 1. Strains used in the present study (Organism, Strain no., Source)	*Arthrobacter nicotianae* CIP 82.107 (ATCC 15236)		Air	8880479
	Arthrobacter oxydans DSM 20119 (ATCC 14358)		Soil	8880479

Directly habitat-related sentence	organisms	relationship	habitats	PMID
	Arthrobacter atrocyaneus CIP 102365 (ATCC 13752)		Soil	8880479
	Arthrobacter aurescens ATCC 13344		Soil	8880479
	Arthrobacter crystallopoietes CIP 102717 (ATCC 15481)		Soil	8880479
	Arthrobacter histidinolovorans ATCC 11442		Soil	8880479
	Arthrobacter pascens ATCC 13346		Soil	8880479
	Arthrobacter ramosus ATCC 13727		Soil	8880479
	Arthrobacter ureafaciens ATCC 7562		Soil	8880479
	Arthrobacter uratoxydans ATCC 21749		Humus soil	8880479
	Arthrobacter protophormiae ATCC 19271		Protophormia terraenovae	8880479
	Arthrobacter cumminsii DMMZ 445 (DSM 10493T)c		Urine	8880479
	Arthrobacter cumminsii DMMZ 483 (DSM 10494)		Urine	8880479
	Arthrobacter cumminsii DMMZ 537		skin infection	8880479
	Arthrobacter woluwensis CUL 1808 (DSM 10495T)		Blood culture	8880479
	Arthrobacter sp. LCDC 92-0385		Blood culture	8880479
	Arthrobacter sp. LCDC 92-0394		Blood culture	8880479
	Arthrobacter sp. LCDC 92-0600		Blood culture	8880479
	Arthrobacter sp. DMMZ 1369		Vaginitis	8880479
Streptomycetes are unique among soil bacteria are unique among soil bacteria	Streptomycetes	are	soil bacteria	12692562
Streptomyces is a genus of Gram-positive bacteria that grows in soil, marshes, and coastal marine habitats	Streptomyces	grows in	soil, marshes, and coastal marine habitats	11572948
Streptomyces avermitilis is a soil bacterium	*Streptomyces avermitilis*	is	soil bacterium	11572948
A. tumefaciens C58, isolated from a cherry tree (Prunus) tumor	*Agrobacterium tumefaciens* C58	isolated from	cherry tree	11743194

Directly habitat-related sentence	**organisms**	**relationship**	**habitats**	**PMID**
Saccharopolyspora erythraea is used for the industrial-scale production of the antibiotic erythromycin A, derivatives of which play a vital role in medicine. The sequenced chromosome of this soil bacterium comprises…	*Saccharopolyspor erythraea*	is	soil bacterium	17369815
This strain was collected from the Sargasso Sea at a depth of 10 meters	*Erythrobacter litoralis* HTCC2594	collected from	sea	NCBI site
This strain (3937j) is wild-type and was isolated from Saintpaulia plants.	*Erwinia chrysanthemia* strain 3937	isolated from	plants	NCBI site
We have isolated 287 strains of Gram-positive bacteria from various soil samples in Belarus. Among these, 55 were identified as *B. subtilis* on the basis of their sensitivity to specific bacteriophages [AR1, AR3, AR9, 0105, and SP01 (Kozlowski and Prozorov, 1981)].	*Bacillus subtilis*	isolated from	soil sample	12584001

Appendix C

Examples of organism-habitat extraction results from the text-mining system developed in this project

The text-mining system processed 9,265 full text documents from the Open Access subset of PubMed Central available in 2007. The results presented in this table are a summary of the organism-habitat pairs extracted by the text mining tool. Only pairs with a valid organism name are included in this table. Extracted pairs with partial or incorrect organism names, such 'A.' or 'fumigatus' were excluded; these incorrect names occurred due to an incorrect extraction by text mining system. The work presented in this table was performed by text-mining experts at the National Centre for Text Mining (NaCTeM[1]) using the system jointly developed and trained by the author and the text-mining experts. Several non-habitat associated terms such as 'process', 'test', and 'helix' were extracted as 'habitats' of organisms, indicating that the system produced a number of false positive results for extracting habitat terms. Most of microorganism names were picked up by the system, however, partial names were sometimes annotated as an organism name, indicating a high degree of sensitivity of the system. An on-going work is being done in order to improve the performance of the test-mining system.

organism entity	habitat entity
A. butzleri	human, animals, humans, marsh, food, water, animal, avian
A. fumigatus	process, mice, humans, macrophage, mouse, human, murine, mammalian, lung, nasal, animal, lungs, udder, airway, solid organ, marrow, toxocara canis
A. halophilus	marine
A. nidulans	human
A. pernix	human

[1] http://www.nactem.ac.uk/, accessed 10th December 2010

organism entity	**habitat entity**
A. phagocytophilum	human, process, tick, ixodes ricinus, vertebrates, test, humans
A. pleuropneumoniae	lung, blood, pigs, animals, liver tissue, tracheobronchial lymph node
A. tumefaciens	plants, process
A. versicolor	feces, larvae, av
A. vinelandii	soil, process, mops
Acropora millepora	animal, animals, nematostella vectensis, dugesia japonica, cnidarians, hydra magnipapillata, mammalian
Acropora palmata	corals
Actinobacillus pleuropneumoniae	swine, respiratory tract, mice, wild boars, domestic pigs, sus scrofa domestica, pigs
Actinobacteria	plants, feces
Actinomyces naeslundii	oral
Actinomyces spp.	oral
Aeromonas salmonicida	salmo salar, oryzias latipes, liver, salmonid, fish, rainbow trout
Aeropyrum pernix	idea
Agrobacterium tumefaciens	plants, soil, crown
Anabaena sp.	nitzschia, human
Anaplasma marginale	drosophila melanogaster, brugia malayi, tick, babesia bovis, bigemina, milk
Anaplasma phagocytophilum	blood, human, ruminants, equine, canine
Archaeoglobus fulgidus	mammals, plants, insects, vertebrates, human
Aspergillus fumigatus	human, respiratory tract, humans, skin, dermatophagoides pteronyssinus, cat, hair, horse, cockroach, virginia, milk, monocytes
Aspergillus nidulans	soil
B. abortus	bovine, canine, porcine, rodent, test, mammal
B. adolescentis	hand, human, swine, faeces, faecal
B. anthracis	cutaneous, lymph nodes, a mouse, anthrax, livestock, humans, food, mouse, feces, mice, lungs, peyer’s patches, human, mammals, guinea pig, animal, pulmonary, lung, airways, guinea pigs
B. bovis	eimeria tenella, toxoplasma gondii, mosquito, tick, sporozoite, cattle, arthropod, human, blood, ticks
B. bronchiseptica	animals, bears, dogs, rabbits, mice, pigs, respiratory tracts, trachea, lungs, rabbit
B. burgdorferi	ticks, tick, ixodes scapularis, humans, vertebrate, ixodes ricinus, scapularis, pacificus, persulcatus, blood, skin, mammal, subject, mice, bladder, heart, vertebrates, spleen, dogs, glomerular, human
B. cereus	food, mice, insects, rabbit, soil, nematodes
B. cetaceae	porpoises, dolphins, minke whales, seals
B. fragilis	animals, mice, intestinal, bias
B. henselae	cat, cardiac valve, human, cats, ctenocephalides
B. japonicum	plants, root
B. mallei	horses, human, humans, animal, equidae, animals, mammalian, individual, turkey
B. maris	mammal
B. melitensis	mammal
B. neotomae	mammal
B. pertussis	animals, human, subject, blood, humans, macrophage, nasopharyngeal, test, mice, mouse, lungs, lung, alveolar, respiratory tract
B. pseudomallei	soil, water, mammalian, avian, humans, livestock, individual, lung, goat, human, animal
B. stearothermophilus	human
B. subtilis	soil, helix, subject, users, human, mouse, rat liver, simplex, humans
B. suis	mammal, pigs, monocytes, macrophage, gastrointestinal tract, stomach, blood, valencia
B. thermophilum	faeces, ruminants, swine, animal, human
B. thetaiotaomicron	large intestine, human, animals
B. thuringiensis	bears
B. vulgatus	human
Babesia	piroplasmida, apicomplexa, sporozoites, blood, mammalian, tick, larvae
Bacillus anthracis	stem, wood, anthrax, mammals
Bacillus cereus	humans, food
Bacillus pumilus	animal
Bacillus stearothermophilus	human
Bacillus subtilis	idea, test, anthrax, water, radius, cota, drosophila melanogaster
Bacteroidales	murine, mice
Bacteroides fragilis	blood
Bacteroides spp.	wastewater, wastewater treatment plant, fecal
Bacteroides thetaiotaomicron	human, tract, oral
Bacteroides vulgatus	intestinal, human
Bartonella	bias, lice, blood, lymph nodes
Bartonella henselae	cats, heart valve
Beggiatoa sp.	individual, marine
Bifidobacterium	colon, intestinal, human, faeces
Bifidobacterium longum	human
Bifidobacterium thermophilum	faeces
Blochmannia spp.	aphids, ant, tsetse fly, insect
Bordetella bronchiseptica	mammals
Bordetella pertussis	chiron, nasopharyngeal, mammals, humans
Borrelia burgdorferi	ticks, mammals, toxoplasma gondii, ixodes, human, simplex, process

organism entity	**habitat entity**
Branchiostoma	protochordates, vertebrate, asymmetron, epigonichthys, gonads, gonadal, invertebrates, the sea
Branchiostoma floridae	tunicate, vertebrate, cephalochordata, lymnaea stagnalis, snail, cephalochordate, invertebrate, vertebrates, drosophila melanogaster, brain
Brucella abortus	animals, bovine
Brucella suis	animal, balb/c
Buchnera aphidicola	aphids, endosymbiont, bias
Burkholderia cepacia	flies, bassiana, soil, biofilm
Burkholderia mallei	horses
Burkholderia pseudomallei	humans, water
C. abortus	felis, domesticated animal, ruminant
C. acetobutylicum	helix
C. albicans	human, macrophage, peritoneal, mammalian, mammals, biofilm, prominent, nematode, vein, mice, soil, oral, oral, blood, experimental animal models, flies, muscle, intestinal tract, humans, intestinal, murine, animals, thrush, esophagus, tongue, mouth, large intestine, cecum, ileum, fecal, mouse, test, athymic, tract, stomach, kidneys, liver, spleen
C. bovis	mites, bovidae, cervidae, equidae, camelidae, horse, mydaus, chorioptes, mite
C. coli	turkeys
C. crescentus	branch, plants
C. difficile	human, animal, oral
C. gingivalis	oral
C. intestinalis	insect, human, invertebrate, fruit fly, worm, fishes, takifugu rubripes, vertebrate, vertebrates, urochordata, cephalochordata, wasp, mammals, opossum, chicken, xenopus, arthropod, humans
C. jejuni	humans, water, poultry, milk, food, animal, human, rabbit, cattle, wild bird, livestock, chicken, avian, chickens, insect, individual, tail, blood, sheep, farm animals, mammalian, tract, gastrointestinal tract, caecum, ileum, intestinal microflora
C. jejuni RM1221	duck
C. merolae	plants, animals
C. neoformans	human, blood, mouse, cerebral, rabbit, crabs, nematodes, flat worms, phyla, pigeon, animals, process, lung, macrophage, animal, mammalian
C. perfringens	chickens, ileum, caecum, intestinal microflora
C. pneumoniae	human, mouse, blood, respiratory tract, vessel, rabbit, vessels, aorta, coronary arteries, animal, frog, koala, horse, root
C. psittaci 6BC	avian, humans, birds
C. reinhardtii	individual, plants, metazoa
C. savignyi	urochordates, halocynthia roretzi, doliolum nationalis, marine invertebrates
C. tepidum	drosophila, human, murine
C. trachomatis	the rates, homo sapiens
C. vulgaris	brachionus calyciflorus
Campylobacter	humans, human, food, animal, livestock, poultry, blood, birds, faeces, water, plants, faecal, chicken, chickens
Campylobacter jejuni	human, poultry, humans, cattle, sheep, animal, tract, chickens, chicken, intestinal
Campylobacter spp.	humans, poultry, water, faecal, avian
Campylobacteraceae	human
Candida albicans	human, urinary tract, sepsis, mammalian, stem, wood, urogenital tract, blood, gastrointestinal tract
Candidatus Serratia symbiotica	endosymbiont
Candidatus sulfidicus	marine
Caulobacter crescentus	asp, bias, human
Chlamydia pneumoniae	lung
Chlamydia psittaci	ruminants, sheep, goats
Chlamydia trachomatis	humans, mouse, animal, toxoplasma gondii
Chlamydomonas reinhardtii	plants, acetabulum, helicosporidium, simulium, alveolata, mammals
Chlamydophila abortus	ruminants, sheep, goats
Chlamydophila pneumoniae	upper respiratory tract
Chromobacterium violaceum	water, soil
Ciona intestinalis	invertebrate, ascidians, urochordates, halocynthia roretzi, doliolum nationalis, tunicate, human, chicken, fugu, fruit fly, nematode, vertebrate, vertebrates, chordates, mouse, frog, ascidian, sea urchin, drosophila melanogaster, marine, muscle, snail, sea squirt, echinodermata urchin, anopheles gambiae, apis mellifera, homo sapiens, mus musculus
Ciona savignyi	sea urchin, echinoderm, vertebrate
Citrobacter rodentium	mouse, intestinal, murine
Clostridium difficile	oral
Clostridium perfringens	chickens
Clostridium spp.	intestinal, murine, mice
Colwellia maris	marine
Corynebacterium glutamicum	soil
Corynebacterium striatum	blood
Cryptococcus neoformans	human, soil, nematodes, bird, human body, adult, rattus norvegicus, homo sapiens, mus musculus, drosophila melanogaster, ap, murine, donor
Cryptosporidium parvum	toxoplasma gondii, alveolata, mammals, intestinal, humans, fish, spironucleus salmonicida, rumen
Cyanothece sp.	endosymbiont

organism entity	**habitat entity**
Cytophaga hutchinsonii ATCC 33406	human oral, soil
D. discoideum	mouse, soil, mammalian, plants, human, animal, opossum, chicken, xenopus, arthropod
D. geothermalis	water, biofilm
D. japonica	dugesia japonica
D. pulex	pseudorasbora parva, fleas
D. radiodurans	idea, subject, human
Daphnia	salmonids, fish, salmonid, lake, arthropods, moths, zooplankton, peromyscus, drosophila, freshwater, daphnia dentifera, laevis, schistosoma mansoni, invertebrates, platyhelminths, schmidtea mediterranea, lottia, pseudorasbora parva
Daphnia pulex	flatworm, schmidtea mediterranea, dugesia japonica, mollusks, lottia, arthropods, freshwater, flea, schistosoma mansoni, water flea, blood fluke, sea urchin, schistosoma, hand, caenorhabditis briggsae, insects, drosophila melanogaster, apis mellifera, anopheles gambiae, aedes, arthropod
Deinococcus radiodurans	human
Desulfotalea psychrophila	hand
Dictyostelium discoideum	physarum polycephalum, gracilis, plants, human, soil
E. carotovora	tract, fat body, larvae
E. coli	subject, marine, backbone, mouse, human, valencia, helix, lung, sepsis, blood, baboon, baboons, monocytes, macrophage, animals, flies, spodoptera frugiperda, insect, plants, mammalian, drosophila, chelator, a p, rabbit, fish, water, indicator, snow, bovine, food, intestine, soil, insects, humans, sheep, cattle, urinary tract, mice, pili, kidney, abscess, ovaries, heart, stomach, gills, intestinal, beetles, biofilm, h and, gastrointestinal tract, mammals, process, wood, animal model, tail, sea urchin, thymus, simplex, murine, users, hand, individual, beak, shoulder, mops, donors, spot, cat, a mouse, rat liver, avian, poultry, chicken, d are, feces, larvae, brain, cardiac, test, drosophila melanogaster, chickens, caecum
E. coli BL21	human
E. coli K-12	biofilm
E. coli K12	biofilm
E. coli O157 H7	cattle, bovine, gastrointestinal tract, human, the gastrointestinal
E. cuniculi	microsporidia, process, animals, mammals, murine
E. faecium	animal
E. histolytica	human, tract, oral, metazoa, plants, abscess, test, trophozoite, trophozoites, liver, entamoeba moshkovskii, blood, intestinal, bigelowiella natans
E. ruminantium	animals, tick, test, ticks, ovine, brain, capillary, brains, ruminants, sheep, lambs, livestock, blood
E. tenax	maggots, anal, larvae, prominent, water, aquatic, animals, drosophila
Ehrlichia ruminantium	ruminants, ticks, goat, sheep, cattle, amblyomma, tick
Eikenella corrodens	abscess, oral, intestinal, genital, cutaneous
Encephalitozoon	diplomonadida, microsporidia, nosema
Encephalitozoon cuniculi	mammals, drosophila melanogaster, homo
Entamoeba histolytica	human, abscess, intestinal, humans, fish, spironucleus salmonicida, rumen, metazoa, plants
Enterobacter cloacae	disk, test, flies
Enterobacteriaceae	colon
Enterococcus faecalis	human, flies, bassiana, intestine, urinary tract, humans, pigs
Enterococcus faecium	chicken, human
Erwinia carotovora	plants
Erythrobacter	soil, water
Erythrobacter litoralis	marine
Escherichia coli	nematodes, food, mouse, s and, animals, monocytes, sigma, thymus, nematode, flies, brenda, homo sapiens, test, human, rat, a mouse, stem, wood, porcine, user, blood, reclinomonas americana, biofilm, urinary tract, pili, bladder, small bowel, sheep, helix, murine, water, organ, subject, operator, transgenic mice, avian, humans, invertebrate, mammalian, individual, endo, microphage, xenopus laevis
Escherichia coli K-12	test, marine
Escherichia coli O157 H7	human, animal model
Escherichia coli-	blood
F. necrophorum	abscess, throat, portal vein, inferior vena cava
F. novicida	human, chicken
F. oxysporum	test, human
F. prausnitzii	chickens, caecum
F. tularensis	mosquitoes, flies, ticks, animals, rabbits, rodents, beavers, food
Fasciola hepatica	human, sheep, cattle, pig, donkey, rat, peritoneal, rabbit, fluke
Fervidobacterium pennivorans	hot spring
Francisella tularensis	lung, lungs, mice
Frankia spp.	plants
Fusobacterium necrophorum	upper respiratory tract, throat, wren
G. lamblia	diplomonad, trophozoites, blood, organ, diplomonads, retortamonads, spironucleus vortens, fish, prominent, animals, microsporidia, intestinal
Geobacter metallireducens	vertebrate, human
Giardia lamblia	homo sapiens, prominent, intestinal, humans, animals, human, toxoplasma gondii, diplomonads, stramenopiles
H. arsenicoxydans	aquatic, water

organism entity	**habitat entity**
H. influenzae	animals, animal, human, mice, nasopharyngeal, respiratory tract
H. japonica	x. laevis
H. marismortui	water
H. polymorpha	water
H. pylori	individual, cancer, human, gastric, stomach, corpus, antrum, stomach, humans, animals, homo sapiens, branch, idea, cardiac, blood
H. salinarum	plants
Haemophilus ducreyi	hand
Haemophilus influenzae	human, resident, murine, a mouse, respiratory tract, humans, mouse
Hahella chejuensis KCTC 2396	water, coastal
Halobacterium salinarum	plants, insects, vertebrates
Halorhodospira halophila	human skin
Helicobacter mustelae	ferret
Helicobacter pylori	bias, water, layer, human stomach, gastric, food, human, subject, humans
Herbaspirillum spp.	soil, gastrointestinal tract
Herminiimonas arsenicoxydans	aquatic
Idiomarina loihiensis L2TR	water, coastal
K. lactis	human, vertebrate
Klebsiella oxytoca	sepsis, blood
Klebsiella pneumoniae	stem, wood, human, urinary tract, blood, pancreas
Kluyveromyces lactis	homo sapiens, canis familiaris, mus musculus, rattus norvegicus, drosophila melanogaster, anopheles gambiae, pan troglodytes, gallus gallus, human, vertebrates
L. acidophilus	bias, subject, gastrointestinal tract, humans, animals
L. braziliensis	oral, sand fly, mice, humans, cutaneous, sigma, human, lymph nodes, viannia
L. casei	gastrointestinal tract, mice, oral, milk, mouse
L. delbrueckii	food
L. helveticus	milk, oral, mice, small intestine lamina propria, animals
L. infantum	protozoa, balb/c, mice
L. innocua	ovine, human
L. interrogans	animal model, guinea pig, hamster, liver, kidney, guinea pigs, humans, pulmonary, animal models, hamsters, human, rattus norvegicus, rodent
L. ivanovii	indicator, water
L. lactis	food, backbone, mouse, mice, indicator, sheep, tail, gastrointestinal tract, knockout mice, intestinal
L. maculans	mice
L. major	mice, c57bl/6, balb/c, macrophage, animals, knockout mice, nippostrongylus brasiliensis, schistosoma mansoni, ear lobe, cutaneous, euglena, leptomonas, leishmania donovani, test, blood, sandflies, transgenic, right ear, salivary gland, hand, sand flies
L. minor	food
L. monocytogenes	ovine, flies, gastrointestinal tract, gastric, food, animals, fecal, oral, guinea pigs, liver and spleen, jejunum, milk, c57bl/6byj, balb/cbyj, mice, test, c57bl/6, mouse strains, blood, heart, animal
L. paracasei	a mouse
L. plantarum	donors, gastrointestinal tract
L. pneumophila	water, macrophage
Lactobacillus acidophilus	gastrointestinal tract, genital
Lactobacillus casei	human, donors, oral, mouse, bovine, milk, mammals
Lactobacillus delbrueckii	human
Lactobacillus paracasei	animal model, human vaginal
Lactobacillus sp.	chickens
Lactobacillus spp.	mice, intestinal
Lactococcus lactis	food, animal, milk, sheep, plants
Lactococcus lactis subsp. cremoris	faecal, gastrointestinal tract
Legionella pneumophila	water, biofilm, macrophage
Legionella spp.	water, indicator
Leishmania braziliensis	lutzomyia intermedia, sand fly, skin, cutaneous, human
Leishmania infantum	murine, mice
Leishmania major	subject, mice, phlebotomus papatasi, flies, blood, macrophage, leishmania donovani, toxoplasma gondii
Leishmania mexicana	human, murine
Leptonema illini	human
Leptospira interrogans	human, humans, mammals, cattle, dogs, pigs, horses
Leptospira spp.	humans, water
Listeria ivanovii	indicator
Listeria monocytogenes	human, rat, lung, animal, food, faeces
M. agalactiae	ruminant, ruminants
M. agalactiae PG2	goat
M. avium	animals, humans, pigs, birds, water, soil, plants, human, porcine, avian, bird, swine, food, mammals, biofilm, bronchial, mice, s and, respiratory tract, lymph node, cattle, sheep, goats, horses
M. avium subsp.	avian, human, porcine, animals, humans, pigs, birds, isolates, mammals
M. avium subsp. hominissuis	pigs, humans, swine, bird, birds, human, porcine, cattle
M. avium subsp.avium	human, pigs, animals
M. avium subsp.hominissuis	pigs, humans, animals

organism entity	habitat entity
M. bovis	badgers, deer, mammals, cattle, lymph nodes, antelope, animal, faeces, soil, bovine, human, humans, buffaloes, nasal, oral, surface water, water, faecal, buffalo, animals, test
M. brevicollis	marine, multicellular animals
M. californianus	unionidae, mussels, bivalves, fusconaia flava, tapes philippinarum, veneridae, mytilidae, mytilus edulis, geukensia demissa
M. capricolum subsp.	ruminants
M. catarrhalis	human, lung, murine, lungs, mice
M. chelonae	oral, respiratory tract
M. fermentans	ruminants, birds
M. gallisepticum	bird
M. grisea	human, mouse
M. jannaschii	human
M. marinum	skin, blood
M. massiliensis	water, blood
M. microti	animal
M. moriokaense	water, lymph node
M. mycoides subsp. SC	lung, cattle
M. oryzae	recipient
M. pinnipedii	fur seal
M. pulchra	mantella baroni
M. pulmonis	murine, swine, avian, fish
M. pusilla	marine
M. smegmatis	inia
M. sp.	mantella cowani
M. stadtmanae	intestine
M. terrae	soil, water, feet, buffalo
M. tuberculosis	human, blood, mice, guinea pigs, lungs, alveolar, animals, water, mouse, spot, skin, guinea pig, pulmonary, macrophage, spleen, lung, liver, humans, homo sapiens, lymph node, lymph nodes, second pathway, plants, mammals, apicomplexa
M. tuberculosis H37Rv	guinea pigs, mice, spot
M. tuberculosis complex	human, blood, lymph node, lymph nodes, lung, respiratory tract
M. ulcerans	arthropods, insects, humans, aquatic, naucoridae, belostomatidae, hemiptera, mice, blood, salivary glands, skin, insect, salivary gland, human, laboratory mice, a mouse, tails, mouse, snails, plants, tail, insect vector, as, human skin, adult, water, phormia, biofilm
M. vaccae	ap, organ, water, fresh water
M. xenopi	respiratory tract, knee, abscess
Magnaporthe grisea	plant
Mannheimia haemolytica	upper respiratory tract, bovines, lung
Mesorhizobium loti	bias, human
Methanobrevibacter smithii	human
Methanococcus maripaludis	salt marsh
Methanothermobacter marburgensis	bovine, sheep
Methanothermobacter thermautotrophicus	mammalian
Micrococcus luteus	rat liver, simplex, humans
Moraxella catarrhalis	respiratory tract, human, lung, pulmonary, humans, mouse
Mycobacterium avium	pulmonary, oral, human, animals, ruminants, animal, intestines, sheep, subject, intestinal
Mycobacterium avium complex	iris
Mycobacterium avium subsp.	marsh
Mycobacterium bovis	animals, bovine, buffaloes, water, buffalo, nasal, oral
Mycobacterium bovis BCG	macrophage
Mycobacterium gordonae	blood
Mycobacterium leprae	cancer
Mycobacterium massiliense	respiratory tract
Mycobacterium sp.	test
Mycobacterium spp.	water
Mycobacterium tuberculosis	blood, rats, sigma, iris, toxoplasma gondii, human, mice, lymph nodes, cattle, bovine, murine, pulmonary, lung, animal
Mycobacterium tuberculosis H37 RA	mice, tail
Mycobacterium ulcerans	human, skin, insect, bones, humans
Mycoplasma haemofelis	canis
Mycoplasma mycoides subsp.	animal
Mytilus californianus	mussel
N. crassa	animal, human, fruit fly, humans
N. gonorrhoeae	human, backbone, macrophage
N. meningitidis	human, blood, humans, nasopharynx, nose, throat, adult, animals, human nasopharyngeal, individual, heart
Naegleria gruberi	soil, freshwater, protozoa
Natronomonas pharaonis	donor

organism entity	habitat entity
Neisseria gonorrhoeae	human, toxoplasma gondii
Neisseria meningitidis	human, rat
Neisseria sp.	humans, mouse
Nitrococcus mobilis	paulinella, donor
Nitrosococcus oceani	marine
Nostoc punctiforme	plants
O. algarvensis	marine, mouth, anus, worm
O. tauri	plants
Olavius algarvensis	marine, worm, mouth
Ostreococcus tauri	plants
P. acidilactici	asp, plants, fruits, gastrointestinal tract, poultry, ducks, animals, water, faeces
P. aeruginosa	soil, process, mice, intestinal, human, tract, villus, lung, nematodes, plants, gastrointestinal tract, biofilm, blood, water, resident, lungs, respiratory tract, individual, cuff, animal models, humans, lamina, vessels, test, tract, mouse, disc, abscess, second pathway
P. berghei	mosquito, oocysts, sporozoites, salivary glands, blood, marrow, mice, trophozoites, transgenic, human, humans, monocytes, cba mice, cerebral, merozoites, livers, mouse, liver, murine, macrophage
P. chabaudi	rodent, mouse, mice, blood, anopheles stephensi, mosquitoes, mosquito, splenic, rats, individual
P. entomophila	tract, fat body, larvae
P. falciparum	blood, human, mosquito, trophozoites, plasmodium ovale, malariae, microvasculature, rodent, sporozoites, funestus group, swine, schizonts, humans, turbo, murine, mouse, mosquitoes, oocyst, water, soil, process, aquatic, transgenic, mice, animal models, nonhuman primates, monocytes, merozoites, animal model, transgenic mice, rabbit, trophozoite, merozoite, thais, oocysts, plants, left hand, right, rhesus, individual, rabbits, schizont, anopheles, anopheles arabiensis, funestus, eimeria tenella, toxoplasma gondii, tick, sporozoite, primates, apicomplexa, canary, liver, anopheles gambiae, stephensi, dinoflagellate
P. graminis	plasmodiophorales, spongospora subterranea
P. knowlesi	merozoites, human, blood, rhesus
P. luminescens	nematodes, soil, insect, heterorhabditis bacteriophora, rhabditidae, larvae, food
P. marinus	human
P. penetrans	nematodes
P. stutzeri	sludge, water, biofilm
P. syringae	plants, spot
P. syringae tomato	plants
P. vivax	human, donors, blood, individual, mosquito, rodent, schizont, rabbit, schizonts, prominent, resident, test, rabbits, rhesus monkey, mosquitoes, primates, humans, oocysts, merozoite, adult, column
P. vulgaris	plants
P. yoelii	blood, mouse, human, transgenic, rodent, mice, spleen, c57bl/6, donor, spleens, mosquitoes, merozoite, merozoites, sporozoites, murine, liver, livers, lungs, sporozoite
Pantoea agglomerans	synovial, blood, human, faeces
Pasteuria spp.	nematode, nematodes, soil
Pectobacterium atrosepticum	plants
Pediococcus acidilactici	faecal, faeces, human
Pediococcus pentosaceus	food
Penicillium marneffei	human, lymph node, blood, rats, rhizomys sinensis
Peptostreptococcus anaerobius	genital
Perkinsus marinus	oyster, haplosporidium nelsoni
Phaeodactylum tricornutum	odontella, salina, endosymbiont
Photorhabdus luminescens	nematode, insects
Pichia	mammalian, plants, human, bungarus fasciatus, necator americanus, hand
Piscirickettsia salmonis	fish, rainbow trout
Plasmodium berghei	oocyst, sporozoite, mosquitoes, blood, mosquito, oocysts, sporozoites, mice, c57bl/6, rodent, spleen, lymph nodes, peyer’s patches, cba mouse, cerebral
Plasmodium chabaudi	microvasculature, rodent, cerebral, mice, murine, nematode, heligmosomoides polygyrus, blood, sporozoites, anopheles stephensi, schistosoma mansoni, spleen
Plasmodium falciparum	mosquito, oocysts, sporozoites, salivary glands, blood, trophozoites, food, merozoite, humans, human, mosquitoes, anopheles dirus, paramecium, toxoplasma, brenda, homo sapiens, individual, test, soil, anopheles gambiae, oocyst, water, vertebrate, anopheles, gambiae, cancer, sporozoite, donors, toxoplasma gondii, apicomplexa, hand, schistosoma haematobium, rodent, plasmodium ovale, malariae, arabiensis, funestus, merus, mammalian, cattle, capillary, anopheles stephensi, blood
Plasmodium vivax	turkey, humans, mosquito, monocytes, human, monkey, rabbit, blood
Plasmodium yoelii	murine, human, rodent, livers, animal, primate
Porphyromonas gingivalis	humans, human
Porphyromonas gingivalis W83	human oral, soil
Prevotella intermedia	humans
Prochlorococcus	paulinella chromatophora, freshwater, paulinella, marine, branch, subject, column, test
Prochlorococcus marinus	water
Prochlorococcus spp.	paulinella chromatophora
Pseudomonas aeruginosa	throat, oral, humans, mouth, respiratory tract, biofilm, lung, stem, wood, abscesses, skin, foot, human, cancer, water

organism entity	**habitat entity**
Pseudomonas putida	rhizosphere
Pseudomonas spp.	human vaginal, stem
Pseudomonas syringae	plants, animals, biofilm
Pseudomonas syringae pv. tomato	plants
R. africae	tick, amblyomma hebraeum, ruminants
R. felis	human
R. rhipicephali	ticks
R. solani	soil, recipient
Ralstonia eutropha	humans
Ralstonia spp.	soil
Rhizoctonia solani	soil, plants
Rhizopus oryzae	alveolata, mammals, nematode, xiphinema index, earthworm, lumbricus rubellus, acanthoscurria gomesiana, human
Rhodopirellula baltica	column
Rhodothermus marinus	wood, human
Rickettsia aeschlimannii	hyalomma marginatum
Rickettsia conorii	bias, rhipicephalus sanguineus
Rickettsia massiliae	human, ticks, cattle
Rickettsia prowazekii	bias
Rickettsia rickettsii	blood
Rothia dentocariosa	urinary tract
S. acidocaldarius	ape, water
S. auratus	vertebrates, halibut, fish, muscle
S. aureus	human, nasal, biofilm, flies, animal models, skin, food, animal model, soft tissue, animal, sepsis, lung, blood, intestinal, hands, oral, intestinal tract, gastric, fecal, tract, outer ear, ewes, mammary, mammary glands, glands, ovine, udder, sheep, teat, mammary gland, users, urinary tract
S. caeruleus	gonorhynchus, chanos chanos, root
S. cerevisiae	oral, humans, water, human, animals, mouse, plants, vertebrates, mammalian, idea, drosophila, mammals, subject, users, drosophila melanogaster, mice, molluscs, vertebrate, tick, cancer, beer
S. coelicolor	resident, soil
S. devriesei	horses, infundibular, human, animals, control animals, oral, infundibulum, teeth, dental
S. enterica	mice, rabbit, water, cattle, humans, animal, human, chicken, poultry, fish, food, animal model, rabbits, cerebral, mouse, mammals, liver, spleen, lymph nodes
S. enterica serotype Typhimurium	abscesses
S. enteritidis	rat colon, ileum, small intestine, ileal mucosa, gastrointestinal tract, pigs, animals, food, lamb, avian
S. epidermidis	intestinal tract, nematodes, biofilm, animal models, guinea pig, mouse
S. exigua	larvae, plants
S. hyicus	skin
S. japonicum	flatworm, schistosoma mansoni, flatworms, adult, veins, bladder, humans, fluke, blood, flat worm, mammalian, pig, intestinal, helminths, platyhelminths, vertebrate
S. lividans	human, salmon
S. marcescens	drosophila, intestinal, human, flies, tract, oral, process
S. meliloti	plants, axis
S. mutans	water, dental, a p, oral
S. paratyphi	cancer, humans, animals
S. pneumoniae	nasopharyngeal, murine, nasal, lungs, mice, flies, blood, sigma, water, forebrain, lung, h and, respiratory tract, test, nasopharynx, rat, alveolar, rats, the rat, pulmonary, individual, animals, heart
S. pneumoniae D39	lung, lungs, mice
S. pombe	mammalian, drosophila, homo, process, vertebrate, animal, human, mouse, plants, mice, prominent, water, test
S. purpuratus	animal, the sea, x. laevis, fishes, takifugu rubripes, suberites domuncula, hydra, drosophila, platynereis, homo sapiens, human, opossum, chicken, xenopus, arthropod, vertebrate, vertebrates, mammalian
S. pyogenes	blood, animal, sheep, horse, human
S. salmonicida	spironucleus barkhanus, fish, spironucleus, birds, mice, trophozoites, blood, organ, axis, diplomonad, diplomonads, retortamonads
S. solfataricus	human
S. sonnei	hand, food
S. thermophilus	human
S. typhi	humans, human, individual
Saccharomyces cerevisiae	homo sapiens, drosophila melanogaster, human, plants, mus musculus, hands, vertebrate, canis familiaris, rattus norvegicus, anopheles gambiae, pan troglodytes, gallus gallus, humans, mouse, onchocerca volvulus, brenda, bos taurus, fish, xenopus laevis, frog, worm, microsporidia, process, mammalian, prominent, sheep, mosquito, chicken, users, rat, opossum, monodelphis domestica, drosophila, homo, physarum polycephalum, animal, xenopus
Salmonella enterica	poultry, spleen, liver, oral, intestinal, small intestine, mouse, c57bl/6, balb/c, murine, brain, mice, resident, subject, mammals, human, food, animal
Salmonella paratyphi C	humans
Salmonella spp.	animal, process, insects, birds, animals, cattle, bone, poultry, prominent, intestinal, faeces

organism entity	habitat entity
Salmonella typhimurium	plants, animals, poultry
Salmonella typhimurium LT2	lamb
Schizosaccharomyces pombe	homo sapiens, canis familiaris, mus musculus, rattus norvegicus, drosophila melanogaster, anopheles gambiae, pan troglodytes, gallus gallus, human, mammalian, mouse, plants, homo, idea, drosophila, beer
Shewanella	marine
Shigella flexneri	gastrointestinal tract, murine, intestinal
Sinorhizobium meliloti	plants
Solobacterium	blood
Staphylococcus Epidermidis	oral
Staphylococcus aureus	biofilm, human, skin, abscesses, food, flies, bassiana, stem, wood, a car, bone, soft tissue, test, intestinal, hands, prominent, ear skin, ears, ewes, sheep
Staphylococcus epidermidis	intestinal tract, biofilm
Staphylococcus sp.	hair follicles, glands, insect
Streptococcus bovis	sepsis
Streptococcus mutans	human
Streptococcus pneumoniae	blood, prominent, flies, bassiana, adult, human, upper respiratory tract, humans, a mouse, respiratory tract, airway, bronchial, rat
Streptococcus pyogenes	animal model, human
Streptococcus thermophilus	food
Streptomyces hygroscopicus	soil
T. brucei	human, subject, mouse, mice, leishmania sp., crithidia fasciculata, trypanosoma equiperdum, trypanoplasma, trypanoplasma borreli, individual, foregut, salivary glands, crocodiles, mammals, sheep, liver, tail, mammalian, tsetse fly, blood, sand fly, process
T. cruzi	leishmania tarentolae, donovani, fish, horse, trypanosoma equiperdum, human, cat, cardiac, muscle, macrophage, splenic, mice, blood, donors, insect vectors, feces, insects, intestinal, water, rhodnius prolixus, humans, leishmania sp., vertebrate, animals, c57bl/6, brain, cerebral, pulmonary, skin, insect, animal
T. forsythensis	human
T. maritima	mouse, sediment
T. neapolitana	marine
T. nigroviridis	pufferfish, teleost, hand, mammalian, human, vertebrates, fugu, guppy, gasterosteus aculeatus, mus musculus, fish
T. parva	skin, schizont, cattle, bovine, sporozoite, human, sporozoites, eimeria tenella, toxoplasma gondii, mosquito, tick, schizonts
T. thermophila	milk, animals, food, water
T. thermophilus	helix, a p
T. vaginalis	human
T. whipplei	plants, duodenal, gastric, intestinal, oral, human, blood, synovial, lymph node, cardiac valve, skeletal muscle, bias
Thalassiosira pseudonana	odontella, diplomonads, stramenopiles, endosymbiont
Theileria parva	cattle, livestock, blood, babesia bovis, tick
Thermoanaerobacter tengcongensis	hot spring
Thermococcus sp.	marine
Treponema pallidum	genital, blood, human, test
Trichomonas vaginalis	intestinal, humans, fish, spironucleus salmonicida, rumen
Trypanosoma brucei	human, humans, animals, mammalian, tsetse fly, drosophila, mammals, zebu cattle, pigs, rat, mouse, insects, blood, insect
Trypanosoma cruzi	mammalian, tetrahymena pyriformis, human, mouse, rat, drosophila melanogaster, crithidia fasciculata, leishmania sp., trypanoplasma borreli, bears, animal, mice, protozoans, apicomplexa, toxoplasma gondii, triatoma infestans
U. maydis	human
Ustilago maydis	human, mouse, onchocerca volvulus
V. cholerae	milk, rabbit, biofilm, aquatic, donor, urinary tract
V. fischeri	squid, individual, organ
Vibrio anguillarum	fish, rainbow trout
Vibrio cholerae	human, aquatic, coastal, humans, water, murine, intestinal
Vibrio fischeri	marine, squid, organ
Vibrio vulnificus	hand, helix, shellfish, water
Wolbachia pipientis	arthropod, nematode
X. campestris	mammalian, vertebrate, mouse, human
Xylella fastidiosa	plant
Y. frederiksenii	throat
Y. intermedia	disc, wastewater, throat, human
Y. pestis	bias
Y. pseudotuberculosis	humans
Yersinia pestis	biofilm, nematode
Yersinia pseudotuberculosis	mammalian
Yersinia spp.	human

organism entity	**habitat entity**
Z. mobilis	food

Appendix D

List of 75 microorganisms whose protein sequences were included in the Blast all-vs-all and all-vs-Refseq searches.

The list of taxa whose proteomes were included in the construction of mucosa-associated extracytoplasmic protein families. A detailed description of the analysis process is described in Section 6.2.2. Seventy-five known mucosa-thriving microbes from bacteria and microbial eukaryotes were selected. The proteomes were derived from six different bacterial phyla: 5 Actinobacteria, 7 Bacteroidetes, 11 Chlamydiae, 15 Firmicutes, 1 Fusobacteria, 31 Proteobacteria, and 5 different protists.

Organism	**Phylum**	**Relationship**	**Colonisation**	**RefSeq accession**	**Proteome size**
Bacteria					
Bifidobacterium adolescentis ATCC 15703	Actinobacteria	GIT mutualist		NC_008618.1	1631
Bifidobacterium animalis subsp. lactis AD011	Actinobacteria			NC_011835.1	1528
Bifidobacterium longum DJO10A	Actinobacteria			NC_010816.1, NC_004253.1, NC_004252.1	2003
Bifidobacterium longum NCC2705	Actinobacteria			NC_004307.2, NC_004943.1	1729
Bifidobacterium longum subsp. infantis ATCC 15697	Actinobacteria			NC_011593.1	2416
Bacteroides fragilis NCTC 9343	Bacteroidetes	GIT mutualist	ileum and colon	NC_003228.3, NC_006873.1	4231
Bacteroides fragilis YCH46	Bacteroidetes		ileum and colon	NC_006297.1, NC_006347.1	4625
Bacteroides thetaiotaomicron VPI-5482	Bacteroidetes			NC_004663.1, NC_004703.1	4816
Bacteroides vulgatus ATCC 8482	Bacteroidetes		ileum and colon	NC_009614.1	4065
Parabacteroides distasonis ATCC 8503	Bacteroidetes	GIT mutualist		NC_009615.1	3850

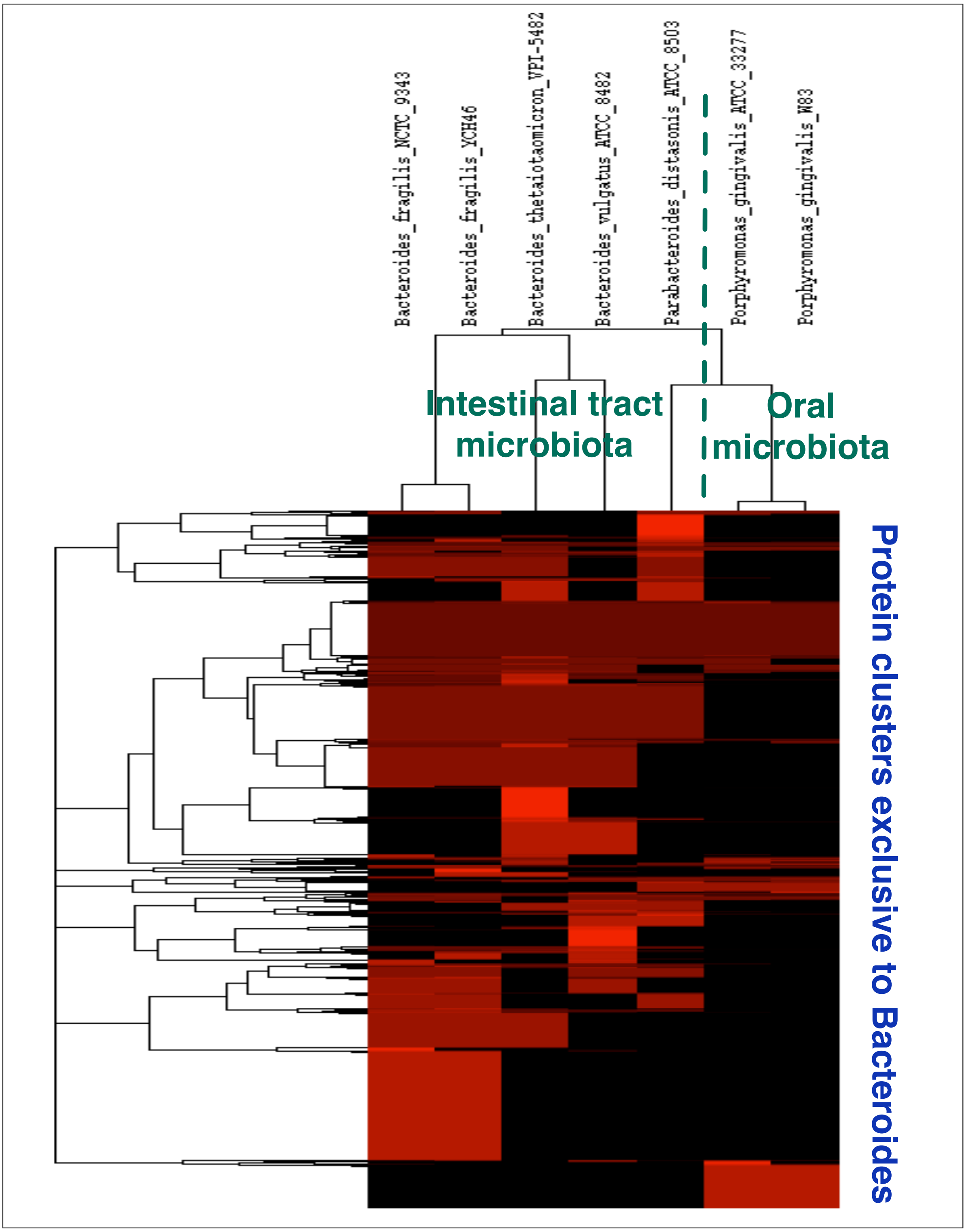
Organism	Phylum	Relationship	Colonisation	RefSeq accession	Proteome size
Porphyromonas gingivalis ATCC 33277	Bacteroidetes		Oral	NC_010729.1	2090
Porphyromonas gingivalis W83	Bacteroidetes		Oral	NC_002950.2	1909
Chlamydia muridarum Nigg	Chlamydiae	Pathogen via mucosal surface		NC_002182.1, NC_002620.2	911
Chlamydia trachomatis 434/Bu	Chlamydiae			NC_010287.1	874
Chlamydia trachomatis A/HAR-13	Chlamydiae			NC_007429.1, NC_007430.1	919
Chlamydia trachomatis D/UW-3/CX	Chlamydiae			NC_000117.1	895
Chlamydophila abortus S26/3	Chlamydiae			NC_004552.2	932
Chlamydophila caviae GPIC	Chlamydiae			NC_003361.3, NC_004720.1	1005
Chlamydophila felis Fe/C-56	Chlamydiae			NC_007900.1, NC_007899.1	1013
Chlamydophila pneumoniae AR39	Chlamydiae			NC_002179.2	1112
Chlamydophila pneumoniae CWL029	Chlamydiae			NC_000922.1	1052
Chlamydophila pneumoniae J138	Chlamydiae			NC_002491.1	1069
Chlamydophila pneumoniae TW-183	Chlamydiae			NC_005043.1	1113
Lactobacillus acidophilus NCFM	Firmicutes	GIT mutualist	ileum and colon	NC_006814.3	1862
Lactobacillus brevis ATCC 367	Firmicutes			NC_008497.1, NC_008499.1	2218
Lactobacillus casei ATCC 334	Firmicutes			NC_008502.1, NC_008526.1	2771
Lactobacillus casei BL23	Firmicutes			NC_010999.1	3044
Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842	Firmicutes			NC_008054.1	1562
Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365	Firmicutes			NC_008529.1	1721
Lactobacillus fermentum IFO 3956	Firmicutes		ileum and colon	NC_010610.1	1843
Lactobacillus gasseri ATCC 33323	Firmicutes			NC_008530.1	1755
Lactobacillus helveticus DPC 4571	Firmicutes			NC_010080.1	1610
Lactobacillus johnsonii NCC 533	Firmicutes			NC_005362.1	1821
Lactobacillus plantarum WCFS1	Firmicutes		colon	NC_004567.1, NC_006375.1- NC_006377.1	3057
Lactobacillus reuteri DSM 20016	Firmicutes			NC_009513.1	1900
Lactobacillus reuteri JCM 1112	Firmicutes			NC_010609.1	1820
Lactobacillus sakei subsp. sakei 23K	Firmicutes			NC_007576.1	1879
Lactobacillus salivarius UCC118	Firmicutes		ileum and colon	NC_007929.1, NC_007930.1, NC_006529.1, NC_006530.1	2013
Fusobacterium nucleatum subsp. nucleatum ATCC 25586	Fusobacteria	pathogen		NC_003454.1	2067
Escherichia coli 536	Proteobacteria-g	pathogen	urinary tract	NC_008253.1	4620
Escherichia coli 55989	Proteobacteria-g	pathogen	GIT	NC_011748.1	4763
Escherichia coli APEC O1	Proteobacteria-g	pathogen	avian lung	NC_009837.1, NC_008563.1, NC_009838.1	4851
Escherichia coli ATCC 8739	Proteobacteria-g			NC_010468.1	4200
Escherichia coli CFT073	Proteobacteria-g	pathogen	urinary tract	NC_004431.1	5339
Escherichia coli E24377A	Proteobacteria-g	pathogen	GIT	NC_009801.1, NC_009786.1- NC_009791.1	5608
Escherichia coli ED1a	Proteobacteria-g			NC_011745.1	4915
Escherichia coli HS	Proteobacteria-g			NC_009800.1	4378
Escherichia coli IAI1	Proteobacteria-g			NC_011741.1	4353

Organism	**Phylum**	**Relationship**	**Colonisation**	**RefSeq accession**	**Proteome size**
Escherichia coli IAI39	Proteobacteria-g	pathogen	GIT	NC_011750.1	4732
Escherichia coli LF82	Proteobacteria-g			NC_011993.1	4312
Escherichia coli O127:H6 str. E2348/69	Proteobacteria-g			NC_011601.1-NC_011603.1	4653
Escherichia coli O157:H7 str. EC4115	Proteobacteria-g			NC_011350.1, NC_011351.1, NC_011353.1	5477
Escherichia coli O157:H7 str. Sakai	Proteobacteria-g			NC_002127.1, NC_002128.1, NC_002695.1	5318
Escherichia coli S88	Proteobacteria-g			NC_011742.1	4696
Escherichia coli SE11	Proteobacteria-g	mutualist	GIT	NC_011416.1, NC_011419.1, NC_011413.1, NC_011411.1, NC_011415.1, NC_011408.1, NC_011407.1	5002
Escherichia coli SMS-3-5	Proteobacteria-g			NC_010498.1, NC_010485.1-NC_010488.1	4913
Escherichia coli str. K-12 substr. DH10B	Proteobacteria-g			NC_010473.1	4126
Escherichia coli str. K-12 substr. MG1655	Proteobacteria-g			NC_000913.2	4131
Escherichia coli str. K-12 substr. W3110	Proteobacteria-g			AC_000091.1	4226
Escherichia coli UMN026	Proteobacteria-g			NC_011749.1, NC_011751.1	4968
Escherichia coli UTI89	Proteobacteria-g			NC_007946.1, NC_007941.1	5166
Pasteurella multocida subsp. multocida str. Pm70	Proteobacteria-g	Pathogen		NC_002663.1	2015
Helicobacter acinonychis str. Sheeba	Proteobacteria-e	Pathogen		NC_008229.1, NC_008230.1	1618
Helicobacter hepaticus ATCC 51449	Proteobacteria-e			NC_004917.1	1875
Helicobacter pylori 26695	Proteobacteria-e		stomach	NC_000915.1	1576
Helicobacter pylori G27	Proteobacteria-e		stomach	NC_011333.1, NC_011334.1	1504
Helicobacter pylori HPAG1	Proteobacteria-e		stomach	NC_008086.1, NC_008087.1	1544
Helicobacter pylori J99	Proteobacteria-e		stomach	NC_000921.1	1489
Helicobacter pylori P12	Proteobacteria-e		stomach	NC_011498.1, NC_011499.1	1578
Helicobacter pylori Shi470	Proteobacteria-e		stomach	NC_010698.2	1569
Microbial eukaryote					
Entamoeba dispar SAW760	Protist-entamoebidae	symbioint		DS547756-DS560014	8812
Entamoeba histolytica HM-1:IMSS	Protist-entamoebidae	Pathogen		DS571145-DS572673	8163
Trichomonas vaginalis G3	Protist-parabasalidea	Pathogen	genitourinary tract	DS113177-DS177945	59518
Cryptosporidium parvum Iowa II	Protist-apicomplexa	Pathogen		NC_006980.1-NC_006987.1	3805
Giardia lamblia ATCC 50803 (Giardia intestinalis ATCC 50803)	Protist-diplomonadida	Pathogen	ileum	CH991761-CH991852	6503

Appendix E

Heatmap dendrogram of extracytoplasmic protein clusters exclusive to *Bacteroides spp*.

The heatmap dendrogram shows the distribution of the protein clusters (vertical axis) among different *Bacteroides species* (horizontal axis) included in the protein clustering analysis. The dendrograms showing the groups of similarity pattern of distribution were constructed by complete linkage hierarchical cluster analysis. Several of these clusters are specific to a particular specie or a group Bacteroides taxa. The distribution shows the variation of genotypic features within the Bacteroides group, suggesting a specific adaptation of the microbes to a variety of selective pressures present in different niches.


Bacteroides_fragilis_NCTC_9343
Bacteroides_fragilis_YCH46
Bacteroides_thetaiotaomicron_VPI-5482
Bacteroides_vulgatus_ATCC_8482
Parabacteroides_distasonis_ATCC_8503
Porphyromonas_gingivalis_ATCC_33277
Porphyromonas_gingivalis_W83
Intestinal tract microbiota
Oral microbiota
Protein clusters exclusive to Bacteroides

Appendix F

A list of mucosa-associated IPR domains without GO term annotation

Given a null hypothesis of no association between an InterPro (IPR) domain and mucosa-thriving microorganisms, a tests based on the hypergeometric distribution yielded significant p-values. Therefore, these IPR domains significantly co-occurred with the mucosa-thriving microorganisms with co-occurrence p-values < 1 $\times 10^{-4}$. The entries were ranked by co-occurrence p-values. The co-occurrence p-value provides the probability of observing the number of organisms within the mucosa-associated microbes with a given protein domain compared to the number of all habitat-classified organisms with that protein domain (reference set). The abundance p-value provides the chance of observing the number of a given domain within the mucosa-associated microbes in comparison to the total number of that domain found in the reference set.

Description	**IPR entry**	**co-occurrence p-value**	**abundance p-value**	**correlation score**
Uracil-DNA glycosylase, active site	IPR018085	3.16×10^{-12}	1.23×10^{-6}	0.30
Prokaryotic chromosome segregation and condensation protein MukF	IPR005582	2.90×10^{-10}	6.24×10^{-10}	0.32
dsDNA mimic, putative	IPR007376	8.75×10^{-10}	6.15×10^{-10}	0.31
Fumarate reductase, subunit C	IPR003510	1.31×10^{-9}	3.33×10^{-9}	0.30
Acid phosphatase (Class B)	IPR005519	2.12×10^{-9}	9.12×10^{-10}	0.29
NLPA lipoprotein	IPR004872	3.30×10^{-9}	5.00×10^{-9}	0.18
Tryptophan/tyrosine permease	IPR018227	7.29×10^{-9}	1.16×10^{-16}	0.23
NGG1p interacting factor 3, NIF3	IPR002678	1.54×10^{-8}	2.02×10^{-4}	0.23
Cyd operon protein YbgE	IPR011846	3.34×10^{-8}	4.78×10^{-8}	0.28
Mannitol repressor	IPR007761	3.34×10^{-8}	2.01×10^{-10}	0.27
Cytochrome b562	IPR009155	3.76×10^{-8}	5.68×10^{-9}	0.28
Spermidine/putrescine import ATP-binding protein, potA	IPR017879	1.42×10^{-7}	2.43×10^{-6}	0.25
YfbU	IPR005587	1.73×10^{-7}	2.46×10^{-7}	0.26
Ribonucleotide reductase Class Ib, NrdI	IPR004465	1.84×10^{-7}	1.05×10^{-7}	0.25
Tellurite resistance methyltransferase, TehB, core	IPR015985	2.23×10^{-7}	3.49×10^{-7}	0.26

Description	**IPR entry**	**co-occurrence p-value**	**abundance p-value**	**corre-lation score**
Phosphomannose isomerase, type I, conserved site	IPR018050	3.34×10^{-7}	2.45×10^{-8}	0.26
Antimicrobial peptide resistance and lipid A acylation PagP	IPR009746	4.20×10^{-7}	1.10×10^{-6}	0.22
Phage shock protein G	IPR014318	4.50×10^{-7}	4.85×10^{-7}	0.26
Porin, general diffusion Gram-negative, conserved site	IPR013793	4.50×10^{-7}	1.70×10^{-20}	0.24
Copper resistance lipoprotein NlpE	IPR007298	4.58×10^{-7}	6.93×10^{-7}	0.25
C4-dicarboxylate anaerobic carrier-like	IPR018385	6.22×10^{-7}	6.93×10^{-13}	0.24
Tryptophan/tryrosine permease, conserved site	IPR013061	6.58×10^{-7}	2.50×10^{-12}	0.25
Glucitol operon activator	IPR009693	8.21×10^{-7}	3.49×10^{-7}	0.24
IlvB leader peptide	IPR012566	1.30×10^{-6}	1.15×10^{-6}	0.25
TfoX, C-terminal	IPR007077	1.36×10^{-6}	2.55×10^{-7}	0.23
WzyE	IPR010691	1.44×10^{-6}	1.47×10^{-6}	0.25
Cytidine and deoxycytidylate deaminase, zinc-binding region	IPR013171	2.01×10^{-6}	3.23×10^{-6}	0.23
Adhesion, bacterial	IPR008966	2.08×10^{-6}	3.29×10^{-97}	0.23
GlpM	IPR009707	2.82×10^{-6}	3.03×10^{-6}	0.24
Ferrous iron transport protein, bacterial	IPR018470	2.83×10^{-6}	1.15×10^{-5}	0.21
Phage shock protein PspD	IPR014321	4.74×10^{-6}	4.41×10^{-6}	0.23
Primosomal replication priB and priC	IPR010890	5.04×10^{-6}	5.50×10^{-6}	0.23
Transcriptional regulator Crl	IPR009986	5.18×10^{-6}	4.51×10^{-6}	0.23
NIL domain	IPR018449	6.57×10^{-6}	8.35×10^{-5}	0.16
Aldose 1-epimerase, conserved site	IPR018052	6.89×10^{-6}	6.31×10^{-6}	0.18
Ionotropic glutamate receptor	IPR001320	8.61×10^{-6}	1.50×10^{-11}	0.21
Phosphotransferase system EIIB/cysteine phosphorylation site	IPR018113	9.07×10^{-6}	3.65×10^{-38}	0.25
DNA damage-inducible protein DinI-like	IPR010391	9.96×10^{-6}	1.58×10^{-10}	0.20
Amino acid transporter, transmembrane	IPR013057	1.08×10^{-5}	7.38×10^{-18}	0.08
YodA	IPR015304	1.08×10^{-5}	5.70×10^{-6}	0.22
Apo-citrate lyase phosphoribosyl-dephospho-CoA transferase	IPR005551	1.35×10^{-5}	1.67×10^{-5}	0.21
Citrate lyase acyl carrier protein CitD	IPR006495	1.35×10^{-5}	2.74×10^{-5}	0.20
Haemolysin expression modulating, HHA	IPR007985	1.38×10^{-5}	2.61×10^{-15}	0.23
Intracellular growth attenuator IgaA	IPR010771	1.38×10^{-5}	1.29×10^{-5}	0.22
LPP motif	IPR006817	1.38×10^{-5}	1.29×10^{-5}	0.22
Secretion monitor	IPR009502	1.38×10^{-5}	1.29×10^{-5}	0.22
SseB	IPR009839	1.38×10^{-5}	1.29×10^{-5}	0.22
Alanine dehydrogenase/pyridine nucleotide transhydrogenase, conserved site-1	IPR008142	1.45×10^{-5}	9.92×10^{-5}	0.17
Glycosyltransferase sugar-binding region containing DXD motif	IPR007577	1.50×10^{-5}	1.96×10^{-11}	0.20
Tetratricopeptide TPR-3	IPR011716	1.88×10^{-5}	6.12×10^{-11}	0.26
Chlamydia cysteine-rich outer membrane protein 6	IPR003506	2.05×10^{-5}	1.77×10^{-5}	0.22
Chlamydia polymorphic membrane, middle domain	IPR011427	2.05×10^{-5}	1.27×10^{-69}	0.21
Cysteine-rich outer membrane protein 3, Chlamydia	IPR003517	2.05×10^{-5}	1.77×10^{-5}	0.22
Histone H1-like Hc1	IPR010886	2.05×10^{-5}	1.77×10^{-5}	0.22
Major outer membrane protein, Chlamydia	IPR000604	2.05×10^{-5}	1.77×10^{-5}	0.22
Phosphotransferase system, EIIC component, type 1	IPR013013	2.27×10^{-5}	1.43×10^{-36}	0.25
Ribonuclease II and R	IPR001900	3.05×10^{-5}	1.74×10^{-4}	0.14
Exonuclease	IPR006055	3.16×10^{-5}	4.02×10^{-6}	0.13
Pili assembly chaperone, conserved site	IPR018046	3.18×10^{-5}	1.02×10^{-32}	0.24
Opacity-associated protein A	IPR007340	3.31×10^{-5}	1.99×10^{-5}	0.17
Opacity-associated protein A, N-terminal	IPR013731	3.08×10^{-5}	8.93×10^{-8}	0.21
Type III secretion sytem,YscO	IPR009929	3.47×10^{-5}	7.19×10^{-5}	0.19
FimH, mannose-binding	IPR015243	3.88×10^{-5}	7.23×10^{-7}	0.21
Mycoplasma MFS transporter	IPR011699	3.88×10^{-5}	7.23×10^{-7}	0.21
Glycosyl transferase, family 8	IPR002495	4.26×10^{-5}	9.27×10^{-20}	0.26
Glycosyl hydrolase family 32, C-terminal	IPR013189	4.60×10^{-5}	1.96×10^{-3}	0.08
HNS-dependent expression A	IPR010486	4.69×10^{-5}	1.40×10^{-5}	0.21
DNA-binding, integrase-type	IPR016177	4.76×10^{-5}	6.84×10^{-6}	0.17
Nicotinate phosphoribosyltransferase-like	IPR015977	4.80×10^{-5}	3.31×10^{-3}	0.12
Exonuclease, RNase T and DNA polymerase III	IPR013520	4.81×10^{-5}	2.86×10^{-6}	0.12

Description	**IPR entry**	**co-occurrence p-value**	**abundance p-value**	**corre-lation score**
YidE/YbjL duplication	IPR006512	5.51×10^{-5}	1.46×10^{-10}	0.22
Anti sigma-E protein RseA, C-terminal	IPR005573	7.15×10^{-5}	8.58×10^{-5}	0.19
FeoC like transcriptional regulator	IPR015102	7.29×10^{-5}	1.23×10^{-4}	0.18
Ribonuclease B, OB region N-terminal	IPR013223	7.35×10^{-5}	2.13×10^{-6}	0.24
CblD like pilus biogenesis initiator	IPR010888	8.09×10^{-5}	6.96×10^{-5}	0.21
Chlamydia 15 kDa cysteine-rich outer membrane	IPR008436	8.09×10^{-5}	6.96×10^{-5}	0.21
Prophage minor tail Z	IPR010633	8.09×10^{-5}	1.23×10^{-9}	0.17
Prophage tail fibre N-terminal	IPR013609	8.09×10^{-5}	8.58×10^{-14}	0.20
ShET2 enterotoxin, N-terminal	IPR012927	8.09×10^{-5}	1.90×10^{-8}	0.20
Citrate lyase ligase, C-terminal	IPR013166	8.33×10^{-5}	9.16×10^{-5}	0.19
Nitrate reductase cytochrome c-type subunit (NapB)	IPR005591	8.43×10^{-5}	1.49×10^{-3}	0.12
Phosphotransferase system, EIIC component, type 2	IPR013014	8.84×10^{-5}	5.71×10^{-26}	0.27
Glucokinase regulatory, conserved site	IPR005486	9.50×10^{-5}	5.27×10^{-6}	0.20
Formate hydrogenlyase maturation HycH	IPR010005	9.83×10^{-5}	4.70×10^{-7}	0.21
Glycosyl hydrolases family 32, N-terminal	IPR013148	9.84×10^{-5}	7.80×10^{-4}	0.08

Appendix G

A list of 64 uncharacterised known protein domains overrepresented in mucosal microorganisms

Given a null hypothesis of no association between an InterPro (IPR) domain and mucosa-thriving microorganisms, a tests based on the hypergeometric distribution yielded significant p-values. Therefore, these IPR domains significantly co-occur with the mucosa-thriving microorganisms with co-occurrence p-values $< 1 \times 10^{-4}$. The entries were ranked by co-occurrence p-values. The co-occurrence p-value provides the probability of observing the number of organisms within the mucosa-associated microbes with a given protein domain compared to the number of all habitat-classified organisms with that protein domain (reference set). The abundance p-value provides the chance of observing the number of a given domain within the mucosa-associated microbes in comparison to the total number of that domain found in the reference set.

Description	**IPR entries**	**co-occurrence p-value**	**abundance p-value**	**correlation score**
Conserved hypothetical protein CHP00022	IPR004375	6.78×10^{-12}	2.29×10^{-20}	0.34
Protein of unknown function DUF1207	IPR009599	2.90×10^{-10}	6.24×10^{-10}	0.32
Conserved hypothetical protein, YtfJ	IPR006513	2.90×10^{-10}	6.24×10^{-10}	0.32
Protein of unknown function DUF441, transmembrame	IPR007382	8.75×10^{-10}	8.50×10^{-13}	0.29
Uncharacterised protein family UPF0242	IPR009623	9.77×10^{-10}	1.87×10^{-9}	0.31
Protein of unknown function DUF328	IPR005583	1.45×10^{-9}	6.39×10^{-16}	0.30
Protein of unknown function DUF1407	IPR010807	6.34×10^{-9}	1.22×10^{-8}	0.29
Protein of unknown function DUF1435	IPR009885	1.72×10^{-8}	1.42×10^{-11}	0.31
Protein of unknown function DUF307	IPR005185	2.38×10^{-8}	5.59×10^{-8}	0.28
Protein of unknown function DUF1144	IPR010574	3.34×10^{-8}	4.78×10^{-8}	0.28
Protein of unknown function DUF1100, hydrolase-like	IPR010520	7.27×10^{-8}	1.55×10^{-9}	0.27
Protein of unknown function DUF462	IPR007411	1.38×10^{-7}	1.58×10^{-7}	0.27

Description	IPR entries	co-occurrence p-value	abundance p-value	correlation score
Protein of unknown function DUF1198	IPR009587	4.20×10^{-7}	4.70×10^{-7}	0.26
Uncharacterised protein family UPF0137	IPR005350	4.50×10^{-7}	1.70×10^{-20}	0.24
Protein of unknown function DUF582	IPR007606	4.50×10^{-7}	4.85×10^{-7}	0.26
Uncharacterised protein family UPF0253	IPR009624	4.50×10^{-7}	1.38×10^{-6}	0.23
Protein of unknown function DUF1131	IPR010938	6.58×10^{-7}	2.50×10^{-12}	0.25
Protein of unknown function DUF30/31	IPR002414	6.73×10^{-7}	1.28×10^{-6}	0.19
Uncharacterized lipoprotein	IPR005619	1.36×10^{-6}	2.55×10^{-7}	0.23
Uncharacterised protein family UPF0259	IPR009627	1.41×10^{-6}	4.17×10^{-7}	0.24
Conserved hypothetical protein CHP00743	IPR005272	1.44×10^{-6}	1.47×10^{-6}	0.25
Protein of unknown function DUF1494	IPR009968	1.66×10^{-6}	2.18×10^{-6}	0.24
Uncharacterised conserved protein UCP006287	IPR008249	2.28×10^{-6}	2.54×10^{-6}	0.24
Protein of unknown function DUF1983	IPR015406	2.92×10^{-6}	2.62×10^{-6}	0.24
Uncharacterised protein family UPF0174	IPR005367	3.41×10^{-6}	3.54×10^{-6}	0.23
Protein of unknown function DUF1137	IPR010564	3.41×10^{-6}	4.36×10^{-7}	0.24
Protein of unknown function DUF1454	IPR009918	3.99×10^{-6}	5.66×10^{-6}	0.23
Protein of unknown function DUF945, bacterial	IPR010352	4.13×10^{-6}	4.94×10^{-10}	0.20
Protein of unknown function DUF419	IPR007351	4.36×10^{-6}	9.81×10^{-13}	0.23
Protein of unknown function DUF720	IPR007966	4.74×10^{-6}	4.41×10^{-6}	0.23
Protein of unknown function DUF1418	IPR010815	4.74×10^{-6}	5.34×10^{-11}	0.15
Protein of unknown function DUF1040	IPR009383	5.04×10^{-6}	5.50×10^{-6}	0.23
Protein of unknown function SprT	IPR006640	6.61×10^{-6}	5.68×10^{-6}	0.22
Uncharacterised protein family UPF0029, N-terminal	IPR001498	8.61×10^{-6}	1.50×10^{-11}	0.21
Protein of unknown function DUF1440	IPR009898	9.96×10^{-6}	1.58×10^{-10}	0.20
Protein of unknown function DUF1706	IPR012550	1.35×10^{-5}	1.67×10^{-5}	0.21
Protein of unknown function DUF413	IPR007335	1.38×10^{-5}	1.29×10^{-5}	0.22
Protein of unknown function DUF496	IPR007458	1.38×10^{-5}	1.29×10^{-5}	0.22
Protein of unknown function DUF1047	IPR009390	1.38×10^{-5}	1.29×10^{-5}	0.22
Protein of unknown function DUF1090	IPR009468	1.38×10^{-5}	1.58×10^{-7}	0.21
Protein of unknown function DUF1480	IPR009950	1.38×10^{-5}	1.29×10^{-5}	0.22
Protein of unknown function DUF414	IPR007336	1.43×10^{-5}	1.11×10^{-6}	0.22
Protein of unknown function DUF687	IPR007787	1.50×10^{-5}	1.96×10^{-11}	0.20
Protein of unknown function DUF1364	IPR010774	1.52×10^{-5}	1.40×10^{-5}	0.22
Uncharacterised protein family UPF0181	IPR005371	2.05×10^{-5}	1.77×10^{-5}	0.22
Protein of unknown function DUF1398	IPR009833	2.05×10^{-5}	1.27×10^{-69}	0.21
Uncharacterised protein family UPF0352	IPR009857	2.05×10^{-5}	1.77×10^{-5}	0.22
Protein of unknown function DUF997	IPR010398	2.05×10^{-5}	1.77×10^{-5}	0.22
Protein of unknown function DUF1471	IPR010854	2.27×10^{-5}	1.43×10^{-36}	0.25
Protein of unknown function DUF535	IPR007488	2.84×10^{-5}	2.20×10^{-5}	0.20
Protein of unknown function DUF533	IPR007486	2.90×10^{-5}	1.03×10^{-3}	0.13
Protein of unknown function DUF388, OB-fold	IPR005220	3.16×10^{-5}	4.02×10^{-6}	0.13
Protein of unknown function DUF991	IPR010393	3.47×10^{-5}	7.19×10^{-5}	0.19
Protein of unknown function DUF986	IPR009328	4.67×10^{-5}	5.76×10^{-5}	0.19
Protein of unknown function DUF1496	IPR009971	4.69×10^{-5}	1.40×10^{-5}	0.21
Uncharacterised protein family UPF0208	IPR007334	5.85×10^{-5}	2.12×10^{-4}	0.17
Protein of unknown function DUF488	IPR007438	6.17×10^{-5}	3.81×10^{-11}	0.17
Protein of unknown function DUF1212	IPR010619	6.17×10^{-5}	9.91×10^{-6}	0.20
Protein of unknown function DUF1347	IPR010764	8.09×10^{-5}	6.96×10^{-5}	0.21
Protein of unknown function DUF1375	IPR010780	8.09×10^{-5}	1.90×10^{-8}	0.20
Protein of unknown function DUF1547	IPR011443	8.33×10^{-5}	9.16×10^{-5}	0.19
Protein of unknown function DUF1481	IPR010858	8.84×10^{-5}	5.71×10^{-26}	0.27
Protein of unknown function DUF1422	IPR009867	9.36×10^{-5}	8.82×10^{-5}	0.19
Protein of unknown function DUF1434	IPR009883	9.83×10^{-5}	4.70×10^{-7}	0.21

Appendix H

Mucosa-associated protein domains located on extracytoplasmic proteins

The mucosa-associated domains identified using co-occurrence p-value cut-off of $< 1 \times 10^{-4}$. A domain was indicated as located on extracytoplasmic protein if more than 50% of all proteins (that were included in this study) carrying that domain were predicted as extracytoplasmic protein by our sequence analysis pipeline. 'Dist.' denotes the distribution of the given domain across superkingdom where A=Archaea, B=Bacteria, E=Eukaryote. 'Class dist.' represents number of taxonomic classification that the domain was annotated. The taxonomic classification system used here are denoted as the sideway headers. 'Total' indicates number of protein sequences predicted to carry the domain. '(%) extprot' represents the proportion of domain-contains sequences that were predicted as extracytoplasmic proteins. Taxonomic classification:

- PROT = Protist: A = Apicomplexa, D = Diplomonadida, E = Entamoebidae, U = Euglenozoa, M = Mycetozoa, and P = Parabasalidea.
- FUN = Fungi: A = Ascomycota, B = Basidomycota, and M = Microsporidia.
- ARC = Archaea: A = Archaeoglobi, H = Halobacteria, Mb = Methanobacteria, Mm = Methanomicrobia, Tc = Thermococci, and Tp = Thermoplasma.
- PRO = Proteobacteria: A = PRO-alpha, B = PRO-beta, D = PRO-delta, E = PRO-epsilon, and G = PRO-gamma.
- Other Bacterial phyla: ACI = acidobacteria, ACT = Actinobacteria, AQU = Aquificae, Bac = Bacteroidetes, CHLA = Chlamydiae, CHLO = Chlorobi, CHLOF = Chloroflexi, CYA = Cyanobacteria, DIC = Dictyoclomi, ELU = Elusimicrobia, FIR = Firmicutes, FUS = Fusobacteria, NIT = Nitrospirae, PLA = Plancotomycetes, SPI = Spirochaetes, TEN = Tenericutes, THEMI = Thermi, THERMO = Thermotogae, VER = verrucomicrobia.

IPR	description	Dist.	Clas dist.	Total	(%) extprot	PROTI	FUN	ARC	ACI	ACT	AQU	BAC	CHLA	CHLO	CHLOF	CYA	DIC	ELU	FIR	FUS	NIT	PLA	PRO	SPI	TEN	THEMI	THEMO	VER
IPR010619	Protein of unknown function DUF1212	ABE	18	433	99.31	M	AB	MbMm		X		X				X				X		X	ABDEG	X				X
IPR006512	YidE/YbjL duplication	AB	16	293	100.00			H	X	X		X			X	X			X	X		X	ABDEG				X	X
IPR004445	Sodium/glutamate symporter	AB	15	266	100.00			Mm	X	X		X				X			X	X		X	ABDEG	X			X	
IPR002053	Glycoside hydrolase, family 25	BE	15	404	67.33	EMP	AB			X		X			X	X			X				ADEG	X				
IPR004872	NLPA lipoprotein	AB	14	929	96.66			A		X		X	X						X	X			ABDEG	X		X	X	
IPR013014	Phosphotransferase system, EIIC component, type 2	AB	13	978	100.00			H		X		X			X				X	X			ABG	X	X	X	X	
IPR005185	Protein of unknown function DUF307	BE	13	209	100.00	MP	AB			X		X		X		X			X				ABDG					
IPR013011	Phosphotransferase system, EIIB component, type 2	AB	13	1612	62.84			H		X					X		X		X	X			ABG	X	X	X	X	
IPR018385	C4-dicarboxylate anaerobic carrier-like	B	12	398	100.00					X		X						X	X	X			ABEG	X	X		X	
IPR006419	Nicotinamide mononucleotide transporter PnuC	B	12	348	100.00					X		X			X	X			X				ABDEG	X				X
IPR013057	Amino acid transporter, trans-membrane	BE	12	499	100.00	ADEMPU	ABM			X									X				G					
IPR000774	Peptidyl-prolyl cis-trans isomerase, FKBP-type, N-terminal	BE	12	520	70.96	A			X			X	X									X	ABDEG	X				X
IPR001127	Phosphotransferase system, sugar-specific permease EIIA 1 domain	AB	12	840	58.81			Mm	X									X	X	X			ABEG	X	X	X		
IPR001320	Ionotropic glutamate receptor	B	11	425	100.00					X					X	X			X				ABDEG			X	X	
IPR013013	Phosphotransferase system, EIIC component, type 1	B	11	1401	100.00					X								X	X	X			ABEG	X	X	X		
IPR018227	Tryptophan/tyrosine permease	AB	11	484	100.00	E		Tc		X			X			X			X				ABDEG					
IPR018113	Phosphotransferase system EIIB/cysteine phosphorylation site	B	11	1409	97.44					X								X	X	X			ABEG	X	X	X		
IPR001996	Phosphotransferase system, EIIB	B	11	1452	95.94					X								X	X	X			ABEG	X	X	X		
IPR005519	Acid phosphatase (Class B)	BE	11	216	87.96		B			X		X		X		X			X				AEG		X			X
IPR004685	Branched-chain amino acid transport system II carrier protein	B	10	483	100.00					X			X						X	X			ABDEG	X				
IPR008966	Adhesion, bacterial	ABE	10	2333	95.71		A	MbMm		X					X				X			X	BDG					
IPR003501	Phosphotransferase system, lactose/cellobiose-specific IIB subunit	B	9	1186	64.42					X					X		X		X				ABG	X	X			

IPR	description	Dist.	Clas dist.	Total	(%) extprot	PROTI	FUN	ARC	ACI	ACT	AQU	BAC	CHLA	CHLO	CHLOF	CYA	DIC	ELU	FIR	FUS	NIT	PLA	PRO	SPI	TEN	THEMI	THEMO	VER
IPR008142	Alanine dehydrogenase/pyridine nucleotide transhydrogenase, conserved site-1	BE	9	222	59.91		A		X	X		X				X			X				ABG					
IPR000849	GlpT transporter	B	8	337	100.00					X			X						X				ABEG	X				
IPR004668	Anaerobic c4-dicarboxylate membrane transporter	B	8	310	100.00					X		X							X				ABDEG					
IPR005220	Protein of unknown function DUF388, OB-fold	B	8	220	98.64									X		X			X				ABDEG					
IPR018470	Ferrous iron transport protein, bacterial	B	8	88	97.73					X		X								X			ABEG	X				
IPR004740	Nucleoside:H+ symporter	B	7	164	100.00				X	X		X						X				X	AG					
IPR004704	Phosphotransferase system, mannose/fructose/sorbose family IID component	AB	7	473	99.58			McTp										X	X	X			DG					
IPR013338	Lysozyme subfamily 2	B	7	355	65.92					X		X							X				ABG				X	
IPR004703	Phosphotransferase system, galactitol-specific IIC component	B	6	193	100.00					X		X							X				ABG					
IPR004700	Phosphotransferase system, sorbose-specific IIC subunit	AB	6	462	100.00			Tp										X	X	X			DG					
IPR004699	Phosphotransferase system, enzyme II sorbitol-specific factor	B	6	91	98.90					X					X				X				ABG					
IPR011638	Sorbitol phosphotransferase enzyme II, C-terminal	B	6	90	98.89					X					X				X				ABG					
IPR009693	Glucitol operon activator	B	6	88	98.86					X					X				X				ABG					
IPR015304	YodA	B	6	122	96.72					X									X				AEG	X				
IPR009898	Protein of unknown function DUF1440	B	6	93	95.70					X									X				ABEG					
IPR011618	Sorbitol phosphotransferase enzyme II, N-terminal	B	6	98	90.82					X					X				X				ABG					
IPR007333	Putative sugar-specific permease, SgaT/UlaA	B	5	289	100.00					X							X		X				G		X			
IPR010352	Protein of unknown function DUF945, bacterial	B	5	172	98.84																		BDEG			X		
IPR005591	Nitrate reductase cytochrome c-type subunit (NapB)	B	5	178	98.31						X												ABEG					
IPR003192	Porin, LamB type	B	5	191	96.34				X														ABDG					
IPR007298	Copper resistance lipoprotein NlpE	B	5	97	95.88							X											BDG	X				
IPR003418	Fumarate reductase, D subunit	B	4	107	100.00					X													ABG					

IPR	description	Dist.	Clas dist.	Total	(%) extprot	PROTI	FUN	ARC	ACI	ACT	AQU	BAC	CHLA	CHLO	CHLOF	CYA	DIC	ELU	FIR	FUS	NIT	PLA	PRO	SPI	TEN	THEMI	THEMO	VER
IPR003510	Fumarate reductase, subunit C	B	4	106	100.00					X													ABG					
IPR009993	4-alpha-L-fucosyltransferase	B	4	79	100.00																		BEG				X	
IPR010398	Protein of unknown function DUF997	B	4	105	100.00														X	X			DG					
IPR009468	Protein of unknown function DUF1090	B	4	113	97.35																		BDEG					
IPR013012	Phosphotransferase system, EIIB component, type 3	B	4	455	83.74					X									X				G	X				
IPR006513	Conserved hypothetical protein, YtfJ	B	3	113	100.00																		DEG					
IPR013061	Tryptophan/tryrosine permease, conserved site	B	3	205	100.00																		BDG					
IPR007382	Protein of unknown function DUF441, transmembrame	B	3	148	100.00														X				BG					
IPR009707	GlpM	B	3	89	100.00																		DEG					
IPR003898	Bordetella pertussis toxin A	B	3	13	100.00																		BG		X			
IPR008992	Enterotoxin, bacterial	B	3	314	99.36														X				BG					
IPR010486	HNS-dependent expression A	B	3	44	97.73																		ABG					
IPR014453	Inhibitor of vertebrate lysozyme	B	3	81	97.53																		ABG					
IPR010780	Protein of unknown function DUF1375	B	3	172	83.14																		DG	X				
IPR000710	Peptidase S6, IgA endopeptidase	B	3	53	79.25																		BEG					
IPR009599	Protein of unknown function DUF1207	B	3	18	77.78								X										BD					
IPR005968	ABC transporter, thiamine, ATP-binding protein	B	2	116	100.00																		AG					
IPR011846	Cyd operon protein YbgE	B	2	91	100.00																		BG					
IPR013793	Porin, general diffusion Gram-negative, conserved site	B	2	288	100.00																		BG					
IPR018046	Pili assembly chaperone, conserved site	B	2	654	100.00																		BG					
IPR009328	Protein of unknown function DUF986	B	2	79	100.00														X				G					
IPR010574	Protein of unknown function DUF1144	B	2	77	100.00																		AG					
IPR009746	Antimicrobial peptide resistance and lipid A acylation PagP	B	2	78	98.72																		BG					
IPR009435	Acid shock	B	2	66	98.48																		BG					

IPR	description	Dist.	Clas dist.	Total	(%) extprot	PROTI	FUN	ARC	ACI	ACT	AQU	BAC	CHLA	CHLO	CHLOF	CYA	DIC	ELU	FIR	FUS	NIT	PLA	PRO	SPI	TEN	THEMI	THEMO	VER
IPR010858	Protein of unknown function DUF1481	B	2	79	97.47																		AG					
IPR005619	Uncharacterized lipoprotein	B	2	118	94.92																		EG					
IPR009155	Cytochrome b562	B	2	99	92.93																		BG					
IPR010888	CblD like pilus biogenesis initiator	B	2	29	86.21																		BG					
IPR002414	Protein of unknown function DUF30/31	B	2	36	80.56														X						X			
IPR014318	Phage shock protein G	B	1	66	100.00																		G					
IPR012566	IlvB leader peptide	B	1	35	100.00																		G					
IPR010691	WzyE	B	1	73	100.00											X												
IPR010771	Intracellular growth attenuator IgaA	B	1	68	100.00																		G					
IPR006817	LPP motif	B	1	81	100.00																		G					
IPR015014	PhoQ Sensor	B	1	68	100.00																		G					
IPR003506	Chlamydia cysteine-rich outer membrane protein 6	B	1	12	100.00								X															
IPR011427	Chlamydia polymorphic membrane, middle domain	B	1	174	100.00								X															
IPR003517	Cysteine-rich outer membrane protein 3, Chlamydia	B	1	12	100.00								X															
IPR000604	Major outer membrane protein, Chlamydia	B	1	12	100.00								X															
IPR011699	Mycoplasma MFS transporter	B	1	17	100.00																				X			
IPR004596	Cell division inhibitor SulA	B	1	66	100.00																		G					
IPR008436	Chlamydia 15 kDa cysteine-rich outer membrane	B	1	11	100.00								X															
IPR012567	IlvGEDA operon leader peptide	B	1	29	100.00																		G					
IPR009627	Uncharacterised protein family UPF0259	B	1	90	100.00																		G					
IPR009867	Protein of unknown function DUF1422	B	1	82	100.00																		G					
IPR009971	Protein of unknown function DUF1496	B	1	67	100.00																		G					
IPR010938	Protein of unknown function DUF1131	B	1	65	100.00																		G					
IPR010815	Protein of unknown function DUF1418	B	1	64	100.00																		G					
IPR009885	Protein of unknown function DUF1435	B	1	63	100.00																		G					

IPR	description	Dist.	Clas dist.	Total	(%) extprot	PROTI	FUN	ARC	ACI	ACT	AQU	BAC	CHLA	CHLO	CHLOF	CYA	DIC	ELU	FIR	FUS	NIT	PLA	PRO	SPI	TEN	THEMI	THEMO	VER
IPR009587	Protein of unknown function DUF1198	B	1	62	100.00																		G					
IPR010564	Protein of unknown function DUF1137	B	1	12	100.00								X															
IPR010764	Protein of unknown function DUF1347	B	1	12	100.00								X															
IPR009968	Protein of unknown function DUF1494	B	1	12	100.00								X															
IPR011443	Protein of unknown function DUF1547	B	1	12	100.00								X															
IPR009623	Uncharacterised protein family UPF0242	B	1	12	100.00								X															
IPR010794	Maltose operon periplasmic	B	1	83	98.80																		G					
IPR009918	Protein of unknown function DUF1454	B	1	68	98.53																		G					
IPR009883	Protein of unknown function DUF1434	B	1	66	98.48																		G					
IPR010854	Protein of unknown function DUF1471	B	1	601	96.17																		G					
IPR007334	Uncharacterised protein family UPF0208	B	1	119	95.80																		G					
IPR015243	FimH, mannose-binding	B	1	60	95.00																		G					
IPR007787	Protein of unknown function DUF687	B	1	28	85.71								X															
IPR009502	Secretion monitor	B	1	63	76.19																		G					

Appendix I

List of taxa and the presence of M60-like domain and their annotated phenotypic features.

This list were used for the association calculation between the M60-like domain and traits of microorganisms; mucosa-associated and animal host-associated. ‘*’ denotes complete genomes available at the time the analyses were performed. ‘**’ a microorganism is labelled as mucosa-associated if there is an evidence showing that at least one of these statements is true: they grow on or colonise mucous membranes; a mucosal environment is a part of their life cycle; they are pathogenic on or through mucosal surfaces; they were isolated from a mucosa-associated area.

Organism*	**Animal host-associated**	**Mucosa associated****	**M60-like domain**	**Isolation site**	**Body sample site**	**Disease**
Acaryochloris marina MBIC11017	non host-associated	no		Prochloron-dominated colonial ascidian Lissoclinum patella off the tropical coast of the Palau islands		None
Acidiphilium cryptum JF-5	non host-associated	no		Coal mine lake sediment		None
Acidithiobacillus ferrooxidans ATCC 23270	non host-associated	no		Acid; bituminous coal mine effluent		None
Acidithiobacillus ferrooxidans ATCC 53993	non host-associated			derived from the type strain DSM 2705		None
Acidothermus cellulolyticus 11B	non host-associated	no		Acid hot spring in Yellowstone		None
Acidovorax avenae citrulli AAC00-1	non host-associated	no				Bacterial fruit blotch
Acidovorax sp JS42	non host-associated	no		Nitrobenzene-contaminated sediment		None
Acinetobacter baumannii AB0057	host-associated	yes		Bloodstream infection of a patient at Walter Reed Army Medical Center.	Blood	Pneumonia; Nosocomial infection
Acinetobacter baumannii AB307-0294	host-associated			blood of a hospitalized patient in Buffalo; NY	Blood	Nosocomial infection

Organism*	**Animal host-associated**	**Mucosa associated****	**M60-like domain**	**Isolation site**	**Body sample site**	**Disease**
Acinetobacter baumannii ACICU	host-associated	yes		Hospital strain from a clone that caused an outbreak in Rome in 2005		Pneumonia; Nosocomial infection
Acinetobacter baumannii ATCC 17978	host-associated	yes		Patient with meningitis		Septicemia; Meningitis; Pneumonia
Acinetobacter baumannii AYE	host-associated	yes				Pneumonia; Nocosomial infection
Acinetobacter baumannii SDF	host-associated			Body lice collected from homeless people living in France		Pneumonia; Nosocomial infection
Acinetobacter calcoaceticus ADP1	host-associated			Derivatove of BD413		Nosocomial infection
Actinobacillus pleuropneumoniae JL03	host-associated	yes		Lung of a pig from a Chinese commercial pig farm in 2003		Necrotizing pleuropneumonia
Actinobacillus pleuropneumoniae L20	host-associated	yes				Porcine pleuropneumonia
Actinobacillus pleuropneumoniae sv 7 AP76	host-associated	yes				Necrotizing pleuropneumonia
Actinobacillus succinogenes 130Z	host-associated	yes		Bovine rumen		None
Aeromonas hydrophila hydrophila ATCC7966	host-associated	yes		Canned milk from the United States		Septicemia; Food poisoning; Gastroenteritis
Aeromonas salmonicida salmonicida A449	host-associated			brown trout in the Eure river; France by Christian Michel in 1975		Furunculosis
Aeropyrum pernix K1	non host-associated	no		Coastal solfataric thermal vent at Kodakara-Jima Island in Japan in 1993		None
Agrobacterium tumefaciens C58-Cereon	non host-associated	no		Cherry tree tumor		Plant tumors
Agrobacterium vitis S4	non host-associated	no		aerial gall that developed on a two-year-old woody grapevine cane		Crown gall
Akkermansia muciniphila ATCC BAA-835	host-associated	yes	+	Human feces	Gastrointestina tract	None
Alcanivorax borkumensis SK2	non host-associated	no		Seawater sediment sample in the Isle of Borkum; North Sea		None
Aliivibrio salmonicida LFI1238	host-associated			Atlantic cod from Hammerfest; Norway		Hitra disease
Alkalilimnicola ehrlichei MLHE-1	non host-associated	no		bottom water from alkaline; hypersaline Mono Lake; CA		None
Alkaliphilus metalliredigenes QYMF	non host-associated	no		Borax leachate ponds		None
Alkaliphilus oremlandii OhILAs	non host-associated	no		Anoxic sediments from the Ohio River; Pennsylvania		None
Alteromonas macleodii Deep ecotype; DSM 17117	non host-associated	no		Seawater obtained at 3500m depth from the Urania Basin in the Mediterranean Sea		None
Anaerocellum thermophilum Z-1320; DSM 6725	non host-associated	no		Hot spring on the Kamchatka peninsula in Russia		None
Anaeromyxobacter dehalogenans 2CP-1	non host-associated	no		Stream sediment near Lansing; Michigan		None
Anaeromyxobacter dehalogenans 2CP-C	non host-associated	no		tropical soil; Cameroon; 1995		None
Anaplasma marginale Florida	host-associated			erythrocytes from pooled blood samples from naturally infected cattle in Florida	Blood	Anaplasmosis; Bovine anaplasmosis
Anaplasma marginale St. Maries	host-associated			Acutely infected cow from Northern Idaho		Anaplasmosis
Anaplasma phagocytophilum HZ	host-associated			Patient in New York in 1995		Anaplasmosis

Organism*	**Animal host-associated**	**Mucosa associated****	**M60-like domain**	**Isolation site**	**Body sample site**	**Disease**
Anoxybacillus flavithermus WK1; DSM 2641	non host-associated	no		waste water drain at the Wairakei geothermal power station in New Zealand		None
Aquifex aeolicus VF5	non host-associated					None
Archaeoglobus fulgidus VC-16	non host-associated	no		Geothermally heated sea floor at Vulcano Island Italy		None
Arcobacter butzleri RM4018	host-associated	yes		Human patient with gastroenteritis		Gastroenteritis; Bacteremia
Arthrobacter aurescens TC1	non host-associated	no		Atrazine-contaminated soil in South Dakota		None
Arthrobacter chlorophenolicus A6	non host-associated	no		Soil at Fort Collins Colorado		None
Arthrobacter sp. FB24	non host-associated	no		Chromate and xylene enriched soil		None
Aspergillus fumigatus Af293	host-associated					Aspergillosis
Aspergillus nidulans FGSC A26(biA1)	host-associated					Aspergillosis
Azorhizobium caulinodans ORS 571	non host-associated	no		Sesbania rostrata;stem nodules		None
Bacillus amyloliquefaciens FZB42	non host-associated	no		Soil		None
Bacillus anthracis Ames	host-associated		+			None
Bacillus anthracis Ames Ancestor A2084 (0581)	host-associated		+			Anthrax
Bacillus anthracis Sterne	host-associated		+			None
Bacillus cereus AH187 (F4810/72)	host-associated	yes	+	Vomit of a person having eaten cooked rice in London UK	Gastrointestina tract	Food poisoning
Bacillus cereus AH820	host-associated	yes	+	October 1995 in Akershus Norway; from the periodontal pocket of a 76 year old female patient with marginal periodontitis	Oral	Food poisoning
Bacillus cereus ATCC 10987	host-associated	yes	+	Cheese spoilage in Canada		Food poisoning
Bacillus cereus B4264 (2002734361)	host-associated	yes	+	1969 from a case of fatal pneumonia in a male patient	Blood	Pneumonia
Bacillus cereus cytotoxis NVH 391-98	non host-associated		+			None
Bacillus cereus DSM 31; ATCC 14579	host-associated		+			Food poisoning
Bacillus cereus E33L (ZK)	host-associated	yes	+	Swab of a zebra carcass in Ethosha National Park in Namibia in 1996		Food poisoning
Bacillus cereus G9842	host-associated	yes	+	Stool samples from an outbreak that involved three individuals in Nebraska in 1996	Gastrointestina tract	Food poisoning
Bacillus cereus Q1	non host-associated	no	+	deep-subsurface oil reservoir in Daqing oilfield; Northeastern China		None
Bacillus clausii KSM-K16	non host-associated					None
Bacillus halodurans C-125	non host-associated			1977		None
Bacillus licheniformis DSM13 Novozymes	non host-associated					Food poisoning
Bacillus pumilus SAFR-032	non host-associated	no		Spacecraft Assembly Facility at NASA Jet Propulsion Laboratory		None
Bacillus subtilis subtilis 168	non host-associated	no		X-ray irradiated strain in Marburg in 1947		None

Organism*	**Animal host-associated**	**Mucosa associated****	**M60-like domain**	**Isolation site**	**Body sample site**	**Disease**
Bacillus thuringiensis Al Hakam	host-associated	no	+	Severe human tissue necrosis	Skin	Sotto disease
Bacillus thuringiensis konkukian 97-27	host-associated	no	+	Severe human tissue necrosis	Skin	Sotto disease
Bacillus weihenstephanensis KBAB4	non host-associated		+			None
Bacteroides fragilis NCTC 9343	host-associated	yes	+		Gastrointestina tract	Abscesses; Diarrhea
Bacteroides fragilis YCH46	host-associated	yes	+		Gastrointestina tract	Abscesses; Diarrhea
Bacteroides thetaiotaomicron VPI-5482	host-associated	yes	+	Feces of a healthy adult	Gastrointestina tract	Peritonitis
Bacteroides vulgatus ATCC 8482	host-associated	yes	+		Gastrointestina tract	Opportunistic peritoneal disease
Bartonella bacilliformis KC583	host-associated				Blood	Carrion's disease; Oroya fever
Bartonella henselae Houston-1	host-associated					Bacillary angiomatosis
Bartonella quintana Toulouse	host-associated					Trench fever; Bacillary angiomatosis; Endocarditis
Bartonella tribocorum CIP 105476	host-associated			Blood of two wild rats; France	Blood	Bartonellosis
Baumannia cicadellinicola Hc	non host-associated			Red portion of the bacteriome from Homalodisca coagulata adults collected in a lemon orchard		None
Bordetella avium 197N	host-associated	yes		Spontaneous nalidixic; acid-resistant derivative of virulent strain 197 which was isolated from an infected turkey		Respiratory infection; Bordetellosis; Coryza
Bordetella bronchiseptica RB50	host-associated	yes		Rabbit		Respiratory infection
Bordetella parapertussis 12822	host-associated	yes		Infected infant in Germany in 1993		Respiratory infection
Bordetella pertussis Tohama I	host-associated	yes		Patient with whooping cough		Respiratory infection
Bordetella petrii Se-1111R; DSM 12804	non host-associated	no		Anaerobic dechlorinating bioreactor culture enriched from river sediment		None
Borrelia afzelii PKo	host-associated	no		Skin lesion from a Lyme disease patient in Europe in 1993	Skin	Acrodermatitis chronica atrophicans; Lyme disease
Borrelia burgdorferi B31	host-associated	no		Dilutional cloning from the original Lyme-disease tick isolate		Lyme disease
Borrelia burgdorferi ZS7	host-associated					Lyme disease
Borrelia duttonii Ly	host-associated	no		2-year-old girl with tick-borne relapsing fever in Tanzania		Tick-borne relapsing fever
Borrelia garinii PBi; OspA	host-associated			Cerebrospinal fluid of a patient with neuroborreliosis in Germany		Lyme disease
Borrelia hermsii DAH	host-associated	no				Tick-borne relapsing fever
Borrelia recurrentis A1	host-associated	no		Adult patient with louse-borne relapsing fever in Ethiopia		Louse-borne relapsing fever
Borrelia turicatae 91E135	host-associated			Soft tick Ornithodoros turicatae in USA		Tick-borne relapsing fever
Bradyrhizobium japonicum USDA110	non host-associated	no		Soybean nodule in 1957 in Florida USA		None
Bradyrhizobium sp BTAi1	non host-associated	no		Stem nodules of Aeschynomene indica		None
Bradyrhizobium sp ORS278	non host-associated	no		Stem nodule of Aeschynomene sensitiva in Senegal in 1991		None

Organism*	Animal host-associated	Mucosa associated**	M60-like domain	Isolation site	Body sample site	Disease
Brucella abortus bv 1; 9-941	host-associated					Brucellosis
Brucella abortus S19	host-associated			Milk of American Jersey Cattle by Dr. John Buck in 1923		Spontaneous abortion
Brucella canis ATCC 23365	host-associated					Canine brucellosis
Brucella melitensis 16M	host-associated			Infected goat		Brucellosis
Brucella melitensis abortus 2308	host-associated			Standard laboratory strain		Brucellosis
Brucella melitensis bv ovis ATCC25840	host-associated			tissue; animal(sheep)		Genital infection; Brucellosis
Brucella melitensis bv suis 1330	host-associated			Swine isolate		Infectious abortions; Brucellosis; Fever
Brucella suis ATCC 23445	host-associated					Infectious abortions; Brucellosis; Fever
Buchnera aphidicola 5A	non host-associated			Acyrthosiphon pisum		None
Burkholderia ambifaria MC40-6	host-associated			cystic fibrosis patient		Cepacia syndrome
Burkholderia cenocepacia AU 1054	host-associated			Blood of a patient with CF	Blood	Pneumonia
Burkholderia cenocepacia HI2424; BCC1	host-associated			Agricultural soil in upstate NY		Pneumonia
Burkholderia cenocepacia J2315	host-associated			Patient with cystic fibrosis in Edinburgh; UK		Chronic infection; Necrotizing Pneumonia
Burkholderia cenocepacia MC0-3	host-associated			Soil associated with maize roots		Chronic infection; Necrotizing Pneumonia
Burkholderia cepacia 383 (R18194)	host-associated			Forest soil in Trinidad in 1958		Chronic infection; Necrotizing Pneumonia
Burkholderia cepacia AMMD	non host-associated	no		Healthy pea plants in Wisconsin in 1985		None
Burkholderia mallei ATCC 23344	host-associated			Chinese patient in Burma who had glanders in 1944		Glanders; Pneumonia
Burkholderia mallei NCTC 10229	host-associated					Glanders; Pneumonia
Burkholderia mallei NCTC 10247	host-associated					Glanders; Pneumonia
Burkholderia mallei SAVP 1	host-associated					Glanders; Pneumonia
Burkholderia multivorans ATCC 17616	host-associated					Cepacia syndrome
Burkholderia phymatum STM815	non host-associated	no		Root nodule of Machaerium lunatum in French Guiana		None
Burkholderia phytofirmans PsJN	non host-associated	no		Surface-sterilized onion roots		None
Burkholderia pseudomallei 1106a	host-associated					Pneumonia; Bacteremia; Melioidosis
Burkholderia pseudomallei 1710b	host-associated					Pneumonia; Bacteremia; Melioidosis
Burkholderia pseudomallei 668	host-associated					Pneumonia; Bacteremia; Melioidosis
Burkholderia pseudomallei K96243	host-associated			Clinical isolate from Thailand		Melioidosis
Burkholderia thailandensis E264	non host-associated	no		Rice field sample in Thailand		None
Burkholderia vietnamiensis G4 (R1808)	non host-associated	no		wastewater; Pensacola; FL		

Organism*	Animal host-associated	Mucosa associated**	M60-like domain	Isolation site	Body sample site	Disease
Burkholderia xenovorans (fungorum) LB400	non host-associated			PCB-containing landfill near in upper New York		Opportunistic infection
Caldicellulosiruptor saccharolyticus DSM 8903	non host-associated	no		Hot spring in New Zealand		None
Caldivirga maquilingensis IC-167	non host-associated	no		Acidic hot spring in the Philippines		None
Campylobacter concisus 13826	host-associated	yes		Feces of a patient with bloody diarrhea in Denmark	Gastrointestina tract	Periodontosis; Gastroenteritis; Gingivitis; Periodontitis
Campylobacter curvus 525.92	host-associated	yes		Feces of a patient with diarrhea in South Africa	Gastrointestina tract	Gastroenteritis; Periodontal infection
Campylobacter fetus fetus 82-40	host-associated			Blood of a human patient who was having a renal transplant	Blood	Infertility; Meningitis; Septicemia; Bacteremia
Campylobacter hominis ATCC BAA-381	host-associated	no		Feces of a healthy human in 2001	Gastrointestina tract	None
Campylobacter jejuni doylei 269.97	host-associated			Human blood	Blood	Bacteremia
Campylobacter jejuni jejuni 81-176	host-associated	yes		Feces of an 9-year-old girl with diarrhea in Minnesota in 1981	Gastrointestina tract	Gastroenteritis; Diarrhea; Food poisoning
Campylobacter jejuni jejuni 81116	host-associated	yes		Waterborne outbreak in 1982		Food poisoning
Campylobacter jejuni jejuni NCTC 11168	host-associated	yes		Diarrheic patient in 1977	Gastrointestina tract	Diarrhea; Food poisoning
Campylobacter jejuni RM1221	host-associated	yes		Skin of a retail chicken by the Food Safety Research Information Office	Skin	Food poisoning; Diarrhea
Campylobacter lari RM2100	host-associated	yes		Human clinical isolate		Gastroenteritis; Bacteremia; Diarrhea; Food poisoning
Candida albicans SC5314	host-associated	yes				Vaginal infection; Oral infection
Candida glabrata CBS 138	host-associated	yes		Human feces	Gastrointestina tract	Human candidasis
Candidatus Azobacteroides pseudotrichonymphae gv. CFP2	host-associated	no		a single cell of termite (Coptotermes formosanus) gut protist		
Candidatus Desulforudis audaxviator MP104C	non host-associated	no		Fracture water from a borehole at a depth of 2.8 km in a South African gold mine		None
Candidatus Korarchaeum cryptofilum OPF8	non host-associated	no		Enriched cells originate from Obsidian Pool; Yellowstone National Park; Wyoming; USA;		None
Candidatus Methanoregula boonei 6A8	non host-associated	no		Acidic peat bog in New York State		None
Candidatus Methanosphaerula palustris E1-9c	non host-associated	no		Rich minerotrophic fen in central New York State		None
Candidatus Phytoplasma aster yellows witches'-broom AY-WB	non host-associated					Aster yellows; Witches' Broom
Candidatus Phytoplasma australiense	non host-associated	no		diseased Chardonnay grapevines from South Australia		Australian grapevine yellows
Candidatus Phytoplasma mali AT	non host-associated	no		Heidelberg; Germany from a symptomatic apple tree		Appleproliferationdisease
Candidatus Phytoplasma onion yellows OY-M	non host-associated	no		Saga Prefecture Japan in 1982		Onions yellow
Candidatus Protochlamydia amoebophila UWE25	host-associated	yes		Environmental isolate; endoSymbiotic of Acanthamoeba sp.		Pneumonia
Candidatus Vesicomyosocius okutanii HA	non host-associated			Hatsushima island in Sagami Bay in Japan		None

Organism*	Animal host-associated	Mucosa associated**	M60-like domain	Isolation site	Body sample site	Disease
Carboxydothermus hydrogenoformans Z-2901	non host-associated	no		Hot swamp from Kunashir Island; Russia		None
Caulobacter crescentus NA1000	non host-associated			variant of wild-type strain CB15		None
Caulobacter sp K31	non host-associated			Low-oxygen groundwater sample		None
Cellvibrio japonicus Ueda107	non host-associated	no		Field soil in Japan		None
Chlamydia muridarum MoPn / Nigg	host-associated	yes		Normal mice; 1939		Respiratory infection; Bronchitis; Pharyngitis; Pneumonia
Chlamydia pneumoniae TW-183	host-associated	yes		Child's conjunctiva during a trachoma vaccine trial in Taiwan in 1965	Eye	Respiratory infection; Pneumonia; Bronchitis; Pharyngitis
Chlamydia trachomatis A/HAR-13	host-associated	yes		conjunctiva isolate	Eye	Pneumonia; Bronchitis; Pharyngitis
Chlamydia trachomatis D/UW-3/CX (sv D)	host-associated	yes				Respiratory infection; Pharyngitis; Trachoma; Venereal disease; Bronchitis; Heart disease; Pneumonia
Chlamydia trachomatis L2/434/BU	host-associated	yes		bubo		Respiratory infection; Bronchitis; Lymphogranuloma vernerum; Pharyngitis; Pneumonia
Chlamydia trachomatis L2b/UCH-1	host-associated	yes		rectal swab of a 49-yr-old MSM who was HIV positive and Hepatitis C negative	Urogenital tract	Respiratory infection; Proctitis; Bronchitis; Pharyngitis; Pneumonia
Chlamydophila abortus S26/3	host-associated	yes		Enzootic abortion case in sheep		Respiratory infection; Bronchitis; Pharyngitis; Pneumonia
Chlamydophila caviae GPIC	host-associated	yes		Guinea pig conjunctiva		Pneumonia; Respiratory infection; Bronchitis; Pharyngitis
Chlamydophila felis Fe/C-56	host-associated	yes				Rhinitis; Respiratory infection; Bronchitis; Pharyngitis; Pneumonia
Chlamydophila pneumoniae AR39	host-associated	yes		University of Washington student with acute respiratory tract infection in 1983	Airways	Respiratory infection; Pneumonia; Pharyngitis; Multiple sclerosis; Heart disease; Bronchitis; Asthma
Chlamydophila pneumoniae CWL029	host-associated	yes				Bronchitis; Respiratory infection; Pharyngitis; Heart disease; Pneumonia
Chlamydophila pneumoniae J138	host-associated	yes		Pharyngeal mucosa of a 5 year old male patient with acute bronchitis in Japan in 1994		Respiratory infection; Pneumonia; Pharyngitis; Asthma; Heart disease; Bronchitis; Multiple sclerosis
Chlorobium limicola DSM 245	non host-associated	no		Gilroy Hot spring		None
Chlorobium phaeobacteroides BS1	non host-associated	no		From the chemocline of the Black Sea		None
Chlorobium phaeobacteroides DSM 266	non host-associated	no		Anoxic sulfide containing water 19.5 m below surface of meromictic Lake Blankvann in Norway		None

Organism*	Animal host-associated	Mucosa associated**	M60-like domain	Isolation site	Body sample site	Disease
Chlorobium tepidum TLS	non host-associated	no		New Zealand high-sulfide hot spring		None
Chloroflexus aggregans DSM 9485	non host-associated	no		Hot spring in Japan		None
Chloroflexus aurantiacus J-10-fl	non host-associated	no		Hakone hot spring area in Japan		None
Chloroflexus sp. Y-400-fl	non host-associated	no		Alkaline hot spring in Little Long Lake in Wisconsin		None
Chloroherpeton thalassium GB 78; ATCC 35110	non host-associated	no		North East coast of the USA		None
Chromobacterium violaceum ATCC 12472	host-associated					Septicemia; Diarrhea
Chromohalobacter salexigens 1H11; DSM 3043	non host-associated	no		Solar salt facility; Netherlands Antilles		None
Citrobacter koseri CDC 4225-83	host-associated			1983 in Maryland where it caused neonatal meningitis		Neonatal meningitis; Bacteremia; Brain abscesses; Meningoencephalitis
Clavibacter michiganensis michiganensis NCPPB 382	non host-associated	no				Tuber rot; Wilting disease; Ring rot
Clavibacter michiganensis sepedonicus ATCC 33113	non host-associated	no		Infected potato		Tuber rot; Ring rot
Clostridium acetobutylicum ATCC 824D	non host-associated	no		Garden soil in Connecticut in USA in 1924		None
Clostridium botulinum A BoNT/A1 ATCC 19397	host-associated	yes		Laboratory strain probably from foodborne botulism cases in the western US		Botulism
Clostridium botulinum A BoNT/A1 Hall	host-associated	yes		Harvard University in 1947		Botulism
Clostridium botulinum B Eklund 17B	non host-associated	no		Marine sediments taken off the coast of Washington; USA		None
Clostridium botulinum BoNT/A3 Loch Maree	host-associated	yes		Duck liver paste during a botulism outbreak at a hotel in the Scottish highlands in 1922.		Botulism
Clostridium botulinum BoNT/B1 Okra	host-associated	yes		Foodborne botulism incident in the US		Botulism
Clostridium botulinum E3 Alaska E43	host-associated	yes	+	Salmon eggs associated with a foodborne case of botulism in Alaska		Botulism
Clostridium botulinum F Langeland	host-associated	yes	+	Home-prepared liver paste involved in an outbreak of foodborne botulism on the island of Langeland in Denmark in 1958		Botulism
Clostridium botulinum type A - Hall	host-associated	yes				Botulism
Clostridium cellulolyticum H10	non host-associated	no		Decayed grass in compost pile		None
Clostridium difficile 630 (epidemic type X)	host-associated	yes		clincal isolate Switzerland		Peritonitis; Colitis; Diarrhea
Clostridium kluyveri DSM 555	non host-associated	no		Mud of a canal in Delft; The Netherlands		None
Clostridium perfringens 13	host-associated	yes	+	Soil isolate		Necrotizing enterocolitis; Enteritis necroticans; Food poisoning; Gas gangrene
Clostridium perfringens ATCC 13124	host-associated	yes	+			Gas gangrene; Food poisoning; Dysenteria; Enterocolitis; Enterotoxemia

Organism*	**Animal host-associated**	**Mucosa associated****	**M60-like domain**	**Isolation site**	**Body sample site**	**Disease**
Clostridium perfringens SM101	host-associated	yes	+			Gas gangrene; Food poisoning; Dysenteria; Enterocolitis; Enterotoxemia
Clostridium phytofermentans ISDg	non host-associated	no		Forest soil near the Quabbin Reservoir in Massachusetts		None
Clostridium tetani Massachusetts E88	host-associated					Tetanus
Colwellia psychroerythraea 34H	non host-associated	no		Arctic marine sediments		None
Coprothermobacter proteolyticus DSM 5265	non host-associated	no		tannery waste containing cattle manure		None
Corynebacterium diphtheriae gravis NCTC 13129	host-associated	yes		Pharyngeal membrane of a 72-year-old unimmunized UK female with clinical diphtheria acquired during a short Baltic cruise in 1997	Airways	Diphtheria; Respiratory infection
Corynebacterium efficiens YS-314; DSM 44549	non host-associated	no		Soil; Japan; Kanagawa	Skin	None
Corynebacterium glutamicum Nakagawa; ATCC 13032	non host-associated			1957 by S. Kinoshita and colleagues while searching for an efficient glutamate-producer		None
Corynebacterium glutamicum R	non host-associated			Meadow soil in Japan		None
Corynebacterium jeikeium K411	host-associated			Axilla of bone marrow transplant patient		Septicemia; Endocarditis; Meningitis; Nosocomial infection
Corynebacterium urealyticum DSM 7109	host-associated	yes		Bladder stone	Bladder	Urinary tract infection; Cystitis; Pyelitis
Coxiella burnetii CbuG_Q212	host-associated	yes		case of endocarditis		Food poisoning; Q fever
Coxiella burnetii CbuK_Q154	host-associated	yes		patient with endocarditis		Food poisoning; Q fever
Coxiella burnetii Dugway 5J108-111 (7E9-12)	host-associated	yes		Rodents in Utah		Food poisoning; Q fever
Coxiella burnetii Nine Mile phase I / RSA 493	host-associated	yes				Food poisoning; Q fever
Coxiella burnetii RSA 331	host-associated	yes		Blood of an infected patient in northern Italy in 1945	Blood	Food poisoning
Cryptococcus neoformans B-3501A	host-associated					Cryptococcosis
Cryptococcus neoformans JEC 21	host-associated					Cryptococcosis
Cryptosporidium parvum Iowa II	host-associated	yes	+			Diarrhea
Cupriavidus metallidurans CH34	non host-associated	no		sedimentation pond in a zinc factory; Belgium		None
Cupriavidus taiwanensis LMG19424	non host-associated	no		Root nodule of the legume Mimosa pudica in Ping-Tung Taiwan; China		None
Cyanothece sp. BH68; ATCC 51142	non host-associated	no		Intertidal sands near Port Aransas; Gulf of Mexico in Texas		None
Cyanothece sp. PCC 7424	non host-associated	no		Rice fields in Senegal		None
Cyanothece sp. PCC 7425	non host-associated	no		Rice fields in Senegal		None
Cyanothece sp. PCC 8801	non host-associated	no		Rice fields in india and Taiwan		None
Dechloromonas aromatica RCB	non host-associated	no		Potomac River Maryland		None

Organism*	Animal host-associated	Mucosa associated**	M60-like domain	Isolation site	Body sample site	Disease
Dehalococcoides sp BAV1	non host-associated	no		environment in Michigan that could dechlorinate polychlorinated ethenes		None
Dehalococcoides sp CBDB1	non host-associated	no		Anoxic river sediment		None
Deinococcus geothermalis DSM11300	non host-associated	no		Hot spring at Agnano in Naples Italy		None
Delftia acidovorans SPH-1	non host-associated	no		Soil enriched with acetamide in Delft in the Netherlands		None
Desulfatibacillum alkenivorans AK-01	non host-associated	no		Sediment from the Arthur Kill; NJ/NY waterway		None
Desulfitobacterium hafniense DCB-2	non host-associated	no		Municipal sludge; Denmark		None
Desulfitobacterium hafniense Y51	non host-associated	no		Soil contaminated with tetrachloroethene in Japan		None
Desulfococcus oleovorans Hxd3	non host-associated	no		Oil tank		None
Desulfotalea psychrophila LSv54	non host-associated	no		Marine sediments off of the coast of Svalbard		None
Desulfotomaculum reducens MI-1	non host-associated	no		Heavy metal-contaminated sediment collected at the Mare Island Naval Shipyard on the San Francisco Bay		None
Desulfovibrio desulfuricans desulfuricans 27774	host-associated	yes		Rumen of a sheep	Gastrointestina tract	None
Desulfovibrio desulfuricans G20	non host-associated	no		Oil well corrosion site		None
Desulfovibrio vulgaris vulgaris DP4	non host-associated	no		heavy metal-impacted lake sediment		None
Desulfovibrio vulgaris vulgaris Hildenborough	non host-associated	no		Clay soil near Hildenborough in UK in 1946		None
Desulfurococcus kamchatkensis 1221n	non host-associated	no		Sediments of Treshchinny Spring (Uzon Caldera; Kamchatka; Russia)		None
Dichelobacter nodosus VCS1703A	non host-associated	no				Ovine footrot
Dictyoglomus thermophilum H-6-12; ATCC 35947	non host-associated	no		Slightly alkaline Tsuetae Hot spring in Kumamoto Prefecture in Japan		None
Dictyoglomus turgidum DSM 6724	non host-associated	no		Hot spring; Uzon volcano caldera; USSR; Kamchatka		None
Dictyostelium discoideum AX4	host-associated	no				
Dinoroseobacter shibae DFL-12	non host-associated			Marine dinoflagellates from the Bay of Tokyo		None
Ehrlichia canis Jake	host-associated			2-year old dog in North Carolina in 1989		Anemia; Ehrlichiosis
Ehrlichia chaffeensis Arkansas	host-associated			Patient on an army base in Arkansas in 1990		Ehrlichiosis; Human monocytic ehrlichiosis
Ehrlichia ruminantium Gardel	host-associated			Caribbean island of Guadeloupe		Heartwater
Ehrlichia ruminantium Welgevonden	host-associated	no		Infected tick		Heartwater
Elusimicrobium minutum Pei 191	non host-associated			Pachnoda ephippiata		None
Entamoeba dispar SAW760	host-associated	yes	+			
Entamoeba histolytica HM-1:IMSS	host-associated	yes	+			Amoebiasis
Enterobacter sakazakii ATCC BAA-894	host-associated			Powdered milk formula fed to a hospitalized neonate that developed an infection		Septicemia; Meningitis; Necrotizing enterocolitis

Organism*	Animal host-associated	Mucosa associated**	M60-like domain	Isolation site	Body sample site	Disease
Enterobacter sp 638	non host-associated			Populus trichocarpa x deltoides		None
Enterococcus faecalis V583	host-associated	yes			Gastrointestina tract	Urinary infection; Bacteremia; Endocarditis
Erythrobacter litoralis HTCC2594	non host-associated	no		Sargasso Sea at a depth of 10m		None
Escherichia coli 55989	host-associated	yes	+			Diarrhea
Escherichia coli C ATCC 8739	host-associated	yes			Gastrointestina tract	None
Escherichia coli DH10B	host-associated	yes		Common laboratory strain; substrain of K-12. It was derived from DH10 (which was derived from MC1061 which in turn was derived from M182)	Gastrointestina tract	None
Escherichia coli IAI1	host-associated	no	+			Meningitis
Escherichia coli IAI39	host-associated	yes	+			Diarrhea
Escherichia coli K-12; MG1655	host-associated	yes	+	Patient in 1922	Gastrointestina tract	None
Escherichia coli LF82	host-associated	yes		patient with Crohn's disease	Gastrointestina tract	Enterocolitis
Escherichia coli O1:K1:H7 APEC	host-associated	yes		Lesion site of a dead turkey with colibacillosis	Gastrointestina tract	
Escherichia coli O127:H6 E2348/69 (EPEC)	host-associated	yes	+			Diarrhea
Escherichia coli O139:H28 E24377A (ETEC)	host-associated	yes			Gastrointestina tract	Diarrhea
Escherichia coli O157:H7 EC4115	host-associated	yes				Diarrhea; Hemorrhagic colitis
Escherichia coli O157:H7 EDL933 (EHEC)	host-associated	yes		raw hamburger meat implicated in hemorrhagic colitis outbreak	Gastrointestina tract	Hemorrhagic colitis; Enterohaemorrhagic; Hamburger disease
Escherichia coli O157:H7 Sakai (EHEC)	host-associated	yes		Outbreak in 1982 in Sakai Japan	Gastrointestina tract	Hemorrhagic colitis; Enterohaemorrhagic; Food poisoning
Escherichia coli O17:K52:H18 UMN026	host-associated	yes	+	Woman with uncomplicated acute cystitis in Minnesota in 1999		Urinary infection
Escherichia coli O45:K1 S88	host-associated	no		Cerebro-spinal fluid of a late onset neonatal meningitis case in France in 1999	Brain	Meningitis
Escherichia coli O6:K15:H31 536 (UPEC)	host-associated	yes		Patient with acute pyelonephritis	Gastrointestina tract	Pyelonephritis; Urinary infection
Escherichia coli O6:K2:H1 CFT073 (UPEC)	host-associated	yes		blood and urine from a woman with acute pyelonephritis; Baltimore Maryland	Blood	Urinary infection; Cystitis; Pyelonephritis
Escherichia coli O81 ED1a	host-associated	yes		Faeces of a healthy man in 2000 in France	Gastrointestina tract	None
Escherichia coli O9 HS	host-associated	yes		Walter Reed Army Institute of Research in 1978	Gastrointestina tract	None
Escherichia coli SE11	host-associated	yes	+	healthy adult human	Gastrointestina tract	None
Escherichia coli SECEC SMS-3-5	non host-associated	no		Toxic-metal contaminated site Shipyard Creek Charleston South Carolina		None
Escherichia coli UTI89 (UPEC)	host-associated	yes	+	Woman with uncomplicated cystitis	Gastrointestina tract	Diarrhea; Cystitis

Organism*	Animal host-associated	Mucosa associated**	M60-like domain	Isolation site	Body sample site	Disease
Exiguobacterium sibiricum 255-15	non host-associated	no		depth of 43.6 m in the permafrost sediment of the Kolyma Indigirka Lowland		None
Fervidobacterium nodosum Rt17-B1	non host-associated	no		Hot spring in New Zealand		None
Finegoldia magna ATCC 29328	host-associated			abdominal wound	Gastrointestina tract	Opportunistic infection; Endocarditis
Flavobacterium johnsoniae UW101	non host-associated	no		Soil in England		Skin lesions
Flavobacterium psychrophilum JIP02/86	host-associated					Bacterial cold water; Rainbow trout fry syndrome
Francisella philomiragia ATCC 25017	host-associated	yes		Water in the Bear River Refuge in Utah		Septicemia; Pneumonia; Bacteremia
Francisella tularensis holarctica FTA (FTNF002-00)	host-associated			from a previously published clinical case in France involving an immunocompetent 56-year old male with bacteremic F.t. holarctica pneumonia		Tularemia
Francisella tularensis mediasiatica FSC147	host-associated			Gerbil in central Asia		Tularemia
Francisella tularensis OSU18	host-associated					Tularemia
Francisella tularensis SCHU S4	host-associated			Human case of tularemia in 1951		Tularemia
Francisella tularensis tularensis A.II; Wyoming; WY96-3418	host-associated			Human finger wound in 1996	Skin	Plague-like illness
Francisella tularensis tularensis FSC 198	host-associated	yes		Slovakia from a mite		Pneumonia; Septicemia
Frankia alni ACN14a	non host-associated			Green alder growing in Tadoussac Canada		None
Frankia sp CcI3	non host-associated	no		Root nodules of Casuarina cunninghamiana in 1983 at Harvard Forest		None
Frankia sp. Mbj2; EAN1pec	non host-associated			Kettering Research Laboratory in Ohio by M Lalonde in 1978		None
Fusarium graminearum PH-1	non host-associated	no				Head blight
Fusobacterium nucleatum nucleatum ATCC 25586	host-associated	yes		Cervico-facial lesion		Periodontal infection
Geobacillus thermodenitrificans NG80-2	non host-associated	no		Oil reservoir formation water taken at a depth of 2000 m		None
Geobacter bemidjiensis Bem	non host-associated	no		Subsurface sediments collected in Bemidji Minnesotta		None
Geobacter lovleyi SZ	non host-associated	no		Noncontaminated creek sediment in June 2002 near Seoul South Korea		None
Geobacter metallireducens GS-15	non host-associated	no		Potomac river downstream of Washington DC in 1987		None
Geobacter sp FRC-32	non host-associated	no		Uranium-contaminated subsurface at US		None
Geobacter sulfurreducens PCA	non host-associated	no		Surface sediments of a hydrocarbon-contaminated ditch in Norman Oklahoma		None
Geobacter uraniumreducens RF4	non host-associated	no		Uranium bioremediation study site in Rifle Colorado		None
Giardia lamblia (intestinalis) WB; clone C6	host-associated	yes				
Gloeobacter violaceus PCC 7421	non host-associated	no		Calcereous rock in Switzerland		None

Organism*	**Animal host-associated**	**Mucosa associated****	**M60-like domain**	**Isolation site**	**Body sample site**	**Disease**
Gluconacetobacter diazotrophicus PAL5	non host-associated	no		Sugarcane roots in Brazil		None
Gramella forsetii KT0803	non host-associated	no		Concentrated seawater collected from the German Bight in the North Sea		None
Granulibacter bethesdensis CGDNIH1	host-associated			39 year old man with Chronic granulomatous disease	Lymph nodes	Chronic granulomatous
Haemophilus ducreyi 35000HP	host-associated			Human skin;upper arm of an experimentally infected human subject	Skin	Chancroid; Genital ulcer
Haemophilus influenzae NTHi 86-028NP	host-associated	yes		Pediatric patient with otitis media from Columbus Children's Hospital	Ear	Sinusitis; Septicemia; Bronchitis; Meningitis; Otitis media
Haemophilus influenzae NTHi PittEE	host-associated	yes		Middle-ear effusion of a child in Pittsburgh		Chronic bronchitis; Otitis media; Meningitis; Septicemia; Sinusitis
Haemophilus influenzae NTHi PittGG	host-associated	yes		External ear discharge of a spontaneously perforated tympanic membrane of a child in Pittsburgh		Chronic bronchitis; Septicemia; Otitis media; Meningitis; Sinusitis
Haemophilus influenzae Rd (KW20)	host-associated	yes		1890s during an influenza pandemic by Pfeiffer	Airways	Bronchitis; Sinusitis; Septicemia; Otitis; Meningitis
Haemophilus parasuis SH0165	host-associated			isolated from a Glasser's disease outbreak farm	Airways	Glasser's disease
Haemophilus somnus 129PT	host-associated					Thrombotic-meningoencephalitis; Septicemia; Pneumonia; Abortion; Constriction of blood vessels; Arthritis; Myocarditis
Haemophilus somnus 2336	host-associated	yes		Lung of a calf which had pneumonia		Pneumonia; Arthritis; Myocarditis
Hahella chejuensis KCTC 2396	non host-associated	no	+	Marine sediment from Cheju Island; Republic of Korea		None
Haloarcula marismortui ATCC43049	non host-associated	no		Dead Sea		None
Haloquadratum walsbyi HBSQ001; DSM 16790	non host-associated	no		Spanish solar saltern		None
Halorhodospira halophila SL1	non host-associated	no		Salt lake mud		None
Halorubrum lacusprofundi ATCC 49239	non host-associated	no		Deep Lake; Antarctica		None
Halothermothrix orenii H 168	non host-associated	no		Salted lake sediment		None
Helicobacter acinonychis Sheeba	host-associated					Gastric lesions
Helicobacter hepaticus 3B1	host-associated					Liver cancer; Gastric bowel disease; Hepatic inflammation; Hepatitis
Helicobacter pylori 26695	host-associated	yes		Patient in the United Kingdom who had gastritis before 1987	Gastrointestina tract	Ulcer; Gastric inflammation
Helicobacter pylori G27	host-associated	yes			Gastrointestina tract	Ulcer; Gastric inflammation; Gastric Ulcerations
Helicobacter pylori HPAG1	host-associated	yes		Swedish patient with chronic atrophic gastritis	Gastrointestina tract	Gastric inflammation; Ulcer
Helicobacter pylori J99	host-associated	yes		Patient with duodenal ulcer in USA in 1994	Gastrointestina tract	Gastric inflammation; Ulcer

Organism*	Animal host-associated	Mucosa associated**	M60-like domain	Isolation site	Body sample site	Disease
Helicobacter pylori P12	host-associated	yes		clinical isolate obtained from a patient with duodenal ulcer	Gastrointestina tract	Gastric inflammation; Ulcer
Helicobacter pylori Shi470	host-associated	yes		clinical isolate from gastric antrum from Amerindian resident of remote Amazonian village of Shimaa; Peru	Gastrointestina tract	Ulcer; Gastric inflammation
Heliobacterium modesticaldum Ice1	non host-associated	no		hot spring microbial mats and volcanic soil in Iceland		None
Herpetosiphon aurantiacus ATCC 23779	non host-associated	no		Birch Lake in Minnesota		None
Hydrogenobaculum sp. Y04AAS1	non host-associated			Obsidian Pool; Acidic hot spring in Yellowstone National Park		None
Hyperthermus butylicus DSM 5456	non host-associated	no		Sea floor of a solfataric environment at 9m depth off the shore of Sao Miguel Island Azores		None
Ignicoccus hospitalis Kin4/I; DSM 18386	non host-associated	no		Shallow marine hydrothermal system of the Kolbeinsey Ridge; north of Iceland		None
Jannaschia sp. CCS1	non host-associated	no		Water sample taken from the North Sea		None
Janthinobacterium sp. Marseille	non host-associated			Solution used in kidney dialysis		None
Kineococcus radiotolerans SRS30216	non host-associated	no		High-level radioactive waste cell at the Savannah River Site in Aiken of South Carolina in 2002		None
Klebsiella pneumoniae Kp342	host-associated	yes		interior of nitrogen-efficient maize plants		Urinary tract infection; Bacteremia; Pneumonia
Klebsiella pneumoniae MGH78578	host-associated	yes		Patient in 1994	Gastrointestina tract	Urinary tract infection; Bacteremia; Pneumonia
Korebacter versatilis Ellin 345	non host-associated	no		soil of an Australian pasture		None
Lactobacillus acidophilus NCFM	host-associated	yes		Human in 1970	Gastrointestina tract	None
Lactobacillus brevis ATCC 367	non host-associated	no				None
Lactobacillus casei ATCC 334	host-associated	yes			Gastrointestina tract	None
Lactobacillus casei BL23	host-associated	yes			Gastrointestina tract	None
Lactobacillus delbrueckii bulgaricus ATCC 11842	non host-associated	no		Bulgarian yogurt in 1919	Gastrointestina tract	None
Lactobacillus delbrueckii bulgaricus ATCC BAA-365	non host-associated	no		Derived from a French starter culture		None
Lactobacillus fermentum IFO 3956	host-associated			fermented plant material in Japan	Gastrointestina tract	None
Lactobacillus gasseri ATCC 33323	host-associated	yes			Gastrointestina tract	None
Lactobacillus helveticus DPC 4571	non host-associated	no				None
Lactobacillus johnsonii NCC533	non host-associated	yes		Human isolate from the Nestle strain collection	Gastrointestina tract	None
Lactobacillus plantarum WCFS1	host-associated	yes		Human saliva	Oral	None
Lactobacillus sakei sakei 23K	non host-associated	no		French sausage		None
Lactobacillus salivarius salivarius UCC118	host-associated	yes		Human gastrointestinal tract		None
Lactococcus lactis lactis IL1403	non host-associated	no		Cheese starter culture		None

Organism*	Animal host-associated	Mucosa associated**	M60-like domain	Isolation site	Body sample site	Disease
Lawsonia intracellularis PHE/MN1-00	host-associated					Proliferative enteropathy
Legionella pneumophila Corby	host-associated			Human isolate		Legionnaire's disease
Legionella pneumophila Lens	host-associated			major outbreak in France		Legionellosis
Legionella pneumophila Paris	host-associated			endemic in France		Legionellosis; Pneumonia
Legionella pneumophila Philadelphia-1	host-associated			initial event of Legionellosis in Philadelphia		Legionellosis
Leifsonia xyli xyli CTCB07	non host-associated					Ratoon stunting
Leishmania braziliensis MHOM/BR/75M2904	host-associated	no				Leishmaniasis
Leishmania infantum JPCM5 (MCAN/ES/98/LLM-877)	host-associated	no				Leishmaniasis
Leishmania major Friedlin	host-associated	no				Visceral disease; Leishmaniasis; Skin ulcer
Leptospira biflexa Patoc 1 (Ames)	host-associated	no		Stream water and kept in the culture collection at the National Animal Disease Center (NADC); Ames; IA since 1990		Leptospirosis
Leptospira biflexa Patoc 1 (Paris)	host-associated	no		stream water and maintained in the collection of the National Reference Center of Leptospira (Institut Pasteur; Paris; France)		Leptospirosis
Leptospira borgpetersenii hardjobovis JB197	host-associated					Leptospirosis
Leptospira borgpetersenii hardjobovis L550	host-associated	no		Human clinical sample		Leptospirosis
Leptospira interrogans Copenhageni Fiocruz L1-130	host-associated	no		Patient with severe leptospirosis during an epidemic in 1996		Leptospirosis
Leptospira interrogans lai 56601	host-associated					Leptospirosis
Leptothrix cholodnii SP-6	non host-associated			Water and flocculent from an artificial iron seep in Ithaca; NY		None
Leuconostoc citreum KM20	non host-associated	no		Baechu kimchi		None
Listeria innocua Clip11262; rhamnose-negative	non host-associated	no		dairy products (cheese) from Morocco		None
Listeria monocytogenes 4b F2365	host-associated	yes		During an outbreak of listeriosis aming patients with AIDS in California in 1985		Listeriosis; Food poisoning
Listeria monocytogenes EGD-e	host-associated	yes		EGD derivative		Food poisoning; Listeriosis
Listeria monocytogenes HCC23	host-associated	yes		channel catfish	Gastrointestina tract	Food poisoning; Listeriosis
Lysinibacillus sphaericus C3-41	non host-associated	no		Mosquito breeding site in China in 1987		Larvicidal toxin
Macrococcus caseolyticus JCSC5402	non host-associated	no		animal meat in a supermarket		None
Magnaporthe grisea 70-15	non host-associated	no				Rice blast
Magnetococcus sp. MC-1	non host-associated	no		Water from the Pettaquamscutt Estuary in Rhode Island		None
Magnetospirillum magneticum AMB-1	non host-associated	no		Pond water in Tokyo Japan		None

Organism*	Animal host-associated	Mucosa associated**	M60-like domain	Isolation site	Body sample site	Disease
Maricaulis maris MCS10	non host-associated	no		Puget Sound in Washington		None
Marinobacter hydrocarbonoclasticus VT8	non host-associated	no		Mediterranean seawater near a petroleum refinery		None
Mesorhizobium loti MAFF303099	non host-associated			from Lotus corniculatus		None
Mesorhizobium sp. BNC1	non host-associated	no		Mixed-culture enriched from sewage using the chelating agent EDTA		None
Metallosphaera sedula DSM 5348	non host-associated	no		Thermal pond in the Pisciarelli Solfatara in Italy		None
Methanobrevibacter smithii PS	non host-associated	no		Sewage digester in Gainesville Florida		None
Methanocaldococcus jannaschii DSM 2661	non host-associated	no		Deep-sea hydrothermal vent in 1982		None
Methanococcoides burtonii DSM6242	non host-associated	no		Anoxic hypolimnion of Ace Lake Antarctica		None
Methanococcus aeolicus Nankai-3	non host-associated	no		Deep marine sediment from the Nankai Trough off the coast of Japan		None
Methanococcus maripaludis C6	non host-associated	no		Intertidal sediments		None
Methanococcus maripaludis C7	non host-associated	no		Intertidal sediments		None
Methanococcus maripaludis S2	non host-associated	no		Salt marsh sediment near Pawley Island South Carolina		None
Methanococcus vannielii SB	non host-associated	no		San Francisco Bay mud flat		None
Methanocorpusculum labreanum Z	non host-associated	no		Surface sediment from the LaBrea Tar Pits in Los Angeles		None
Methanoculleus marisnigri JR1	non host-associated	no		Sediment from the Black Sea		None
Methanopyrus kandleri AV19	non host-associated	no		Black smoker from the Gulf of California at a depth of 2000m		None
Methanosarcina acetivorans C2A	non host-associated	no		Marine sediment		None
Methanosarcina barkeri Fusaro	non host-associated	no		Mud samples from Lago del Fusaro Lake in Naples Italy		None
Methanosphaera stadtmanae DSM 3091	host-associated	yes		Human feces	Gastrointestina tract	None
Methanospirillum hungateii JF-1	non host-associated	no		Sewage sludge		None
Methanothermobacter thermoautotrophicus Delta H	non host-associated	no		Sewage sludge in 1971 in Urbana Illinois		None
Methylacidiphilum infernorum V4	non host-associated	no		Hell's Gate geothermal area in New Zealand		None
Methylibium petroleiphilum PM1	non host-associated	no		Compost biofilter from a water pollution treatment plant in Los Angeles		None
Methylobacillus flagellatus KT	non host-associated	no		Activated sludge found at the wastewater treatment plant in Moscow Russia		None
Methylobacterium chloromethanicum CM4	non host-associated	no		Soil at a petrochemical factory in Tatarstan Russia		None
Methylobacterium nodulans ORS2060	non host-associated	no		Root nodules from the legume Crotalaria		None
Methylocella silvestris BL2	non host-associated	no		Acidic forest cambisol near Marburg Germany		None
Microcystis aeruginosa NIES-843	host-associated	no		Lake Kasumigaura Ibaraki Japan from Otsuka; Shigeto		Gastroenteritis; Skin irritation; Hepatic inflammation
Moorella thermoacetica ATCC39073	non host-associated	no		Bottom of stagnant ponds		None

Organism*	Animal host-associated	Mucosa associated**	M60-like domain	Isolation site	Body sample site	Disease
Mycobacterium abscessus CIP 104536	host-associated			human knee infection with subcutaneous abscess-like lesions	Skin	Wound infection; Broncho-pulmonary infection; Respiratory infection
Mycobacterium avium 104	host-associated			AIDS patient		Tuberculosis type pulmonary infection; Respiratory infection
Mycobacterium avium paratuberculosis K-10	host-associated			Bovine feces; isolated from a dairy herd; Wisconsin in the 1970's	Gastrointestina tract	Paratuberculosis; Johne's disease; Enteritis
Mycobacterium bovis AF2122/97(spoligotype 9)	host-associated			In 1997 in the UK from a cow suffering necrotic lesions in lung and broncho-mediastinal lymph nodes		Tuberculosis
Mycobacterium bovis BCG Pasteur 1173P2	host-associated					Bovine tuberculosis
Mycobacterium flavenscens PYR-GCK	non host-associated			River sediment		None
Mycobacterium leprae Br4923	host-associated			human skin biopsy in Brazil	Skin	Leprosy
Mycobacterium leprae TN	host-associated			Armadillo in Tamil Nadu India		Hanson's disease; Leprosy
Mycobacterium marinum M; ATCC BAA-535	host-associated			Human patient isolate from Moffett Hospital; University of California; San Francisco in 1992		Tuberculosis
Mycobacterium smegmatis MC2 155	host-associated			Mutant of M smegmatis isolated on 1990		Soft tissue lesions
Mycobacterium sp JLS	non host-associated	no		Creosote-contaminated soil from the Champion International Superfund site in Libby Montana		None
Mycobacterium sp KMS	non host-associated			Creosote-contaminated soil from the Champion International Superfund site in Libby Montana		None
Mycobacterium sp MCS	non host-associated			Creosote-contaminated soil from the Champion International Superfund site in Libby Montana		None
Mycobacterium tuberculosis CDC1551 (Oshkosh)	host-associated	yes		Clothing factory worker from the Kentucky/Tennessee		Tuberculosis
Mycobacterium tuberculosis F11 (ExPEC)	host-associated	yes		Tuberculosis patients during a TB epidemic in the Western Cape of South Africa		Tuberculosis
Mycobacterium tuberculosis H37Ra	host-associated	yes		Original human-lung H37 isolate in 1934	Airways	Tuberculosis
Mycobacterium tuberculosis H37Rv	host-associated	yes		Human-lung H37 isolate in 1934	Airways	Tuberculosis
Mycobacterium ulcerans Agy99	host-associated			ulcerative lesion on the right elbow of a female patient from the Ga district of Ghana in 1999	Skin	Buruli ulcer
Mycobacterium vanbaalenii PYR-1	non host-associated	no		oil-contaminated sediment in Redfish Bay; TX; USA		None
Mycoplasma agalactiae PG2	host-associated					Pneumonia; Arthritis; Mycoplasmosis; Mastitis
Mycoplasma arthritidis 158L3-1	host-associated					Arthritis
Mycoplasma capricolum capricolum California kid ATCC 27343	host-associated					Septicemia; Arthritis; Caprine mycoplasma
Mycoplasma gallisepticum R	host-associated					Respiratory infection

Organism*	Animal host-associated	Mucosa associated**	M60-like domain	Isolation site	Body sample site	Disease
Mycoplasma genitalium G-37	host-associated	yes		Urethral specimen from a male patient with non-gonococcal urethritis	Urogenital tract	Urogenital infection; Non-gonococcal urethritis; Respiratory infection
Mycoplasma hyopneumoniae 232	host-associated					Pneumonia; Porcine enzootic pneumonia
Mycoplasma hyopneumoniae 7448	host-associated	yes		Infected pig in Lindoia do Sul Santa Catarina Brazil		Enzootic pneumonia; Respiratory infection
Mycoplasma mycoides mycoides SC PG1T	host-associated	yes				Pleuropneumonia; Respiratory infection
Mycoplasma penetrans HF-2	host-associated	yes	+			Respiratory infection; Urogenital infection
Mycoplasma pneumoniae M129	host-associated	yes		Patient with atypical pneumonia	Airways	Respiratory infection; Tracheobronchitis; Pneumonia
Mycoplasma pulmonis UAB CTIP	host-associated	yes		Laboratory strain		Genital infection; Respiratory infection
Mycoplasma synoviae 53	host-associated	yes		Broiler breeder pig in Parana Brazil		Respiratory infection
Natranaerobius thermophilus JW/NM-WN-LF	non host-associated	no		Soda lakes of the Wadi An Natrun; Egypt		None
Natronomonas pharaonis DSM 2160; Gabara	non host-associated	no		Lake Gabara Egypt		None
Neisseria gonorrhoeae FA1090	host-associated	yes		Patient with disseminated gonococcal infection		Gonorrhea
Neisseria gonorrhoeae NCCP11945	host-associated	yes		Vaginal smear of a Korean patient	Urogenital tract	Gonorrhea
Neisseria meningitidis 053442	host-associated			2003-2005 outbreak of meningococcal disease in China		Septicemia; Meningitis
Neisseria meningitidis C; FAM18	host-associated			Patient with meningococcal septicemia	Blood	Meningitis; Septicemia
Neisseria meningitidis MC58	host-associated			Case of invasive infection.		Meningitis; Septicemia
Neisseria meningitidis Z2491	host-associated			Cerebrospinal fluid; Gambia in 1983	Brain	Meningitis; Septicemia
Neorickettsia sennetsu Miyayama	host-associated			1953 in Japan causing Sennetsu fever	Blood	Sennetsu fever
Nitratiruptor sp SB155-2	non host-associated	no		30-m-tall sulfide mound in the Iheya North field; Japan		None
Nitrobacter hamburgensis X14	non host-associated	no		Soil		None
Nitrosococcus oceani C-107	non host-associated	no		seawater; North Atlantic		None
Nitrosopumilus maritimus SCM1	non host-associated	no		Salt-water aquarium		None
Nocardia farcinica IFM 10152	host-associated	yes		Bronchus of a 68-year-old male Japanese patient	Airways	Mastitis; Nocardiosis
Nocardioides sp. JS614	non host-associated			soil; Carson; CA; USA		None
Nostoc punctiforme ATCC 29133	non host-associated			gymnosperm cycad Macrozamia sp		None
Novosphingobium aromaticivorans DSM 12444 (F199)	non host-associated	no		410m depth from a borehole sample that was drilled at the Savannah River Site in South Carolina		Death of coral reefs
Oceanobacillus iheyensis HTE831	non host-associated	no		Deep sea mud at 1050m depth from the Iheya ridge near Okinawa Japan in 1998		None

Organism*	**Animal host-associated**	**Mucosa associated****	**M60-like domain**	**Isolation site**	**Body sample site**	**Disease**
Ochrobactrum anthropi ATCC 49188	host-associated					Meningitis
Oenococcus oeni PSU-1	non host-associated	no		Wine; Bob Beelman; Penn State University		None
Opitutus terrae PB90-1	non host-associated	no		Rice paddy soil		None
Orientia tsutsugamushi Boryong	host-associated			Korean patient in Boryong in 1995		Scrub typhus
Orientia tsutsugamushi Ikeda	host-associated			patient in Niigata Prefecture; Japan		Scrub typhus
Parabacteroides distasonis ATCC 8503	host-associated					Opportunistic peritoneal disease
Parvibaculum lavamentivorans DS-1	non host-associated	no		Sewage treatment plant in Germany		None
Pasteurella multocida Pm70 type A	host-associated	yes		Case of cholera in chickens in 1995		Septicemia; Cholera; Pasteurellosis
Pectobacterium atrosepticum SCRI1043	non host-associated					Soft rot
Pelobacter propionicus DSM 2379	non host-associated	no		Creek mud; Germany		None
Pelodictyon luteolum DSM 273	non host-associated	no		Meromictic Lake Polden in Norway		None
Pelotomaculum thermopropionicum SI	non host-associated	no		Thermophilic upflow anaerobic sludge blanket reactor		None
Petrotoga mobilis SJ95t	non host-associated	no		Oil field water; North Sea; Norwegian sector		None
Phenylobacterium zucineum HLK1	host-associated			Human erythroleukemia cell line K562	Blood	Unknown
Photobacterium profundum SS9	non host-associated			2500m depth from the Sulu Trough		None
Photorhabdus luminescens laumondii TT01	host-associated	no		Isolated from the nematode Heterorhabditis bacteriophora on Trinidad and Tobago		Toxemia; Septicemia
Plasmodium falciparum	host-associated					Malaria
Plasmodium falciparum 3D7	host-associated	no				Malaria
Plasmodium vivax Salvador I	host-associated	no		naturally-infected honey bee colony at Weslaco; Texas		Malaria
Polaromonas naphthalenivorans CJ2	non host-associated	no		Naphthalene-contaminated freshwater sediment		None
Polaromonas sp. JS666	non host-associated	no		Sediment contaminated with cis-dichloroethane		None
Porphyromonas gingivalis ATCC 33277	host-associated	yes		human gingival sulcus	Oral	Dental plaque; Periodontal infection
Porphyromonas gingivalis W83	host-associated	yes		Human clinical specimen from abscess; Bonn; Germany	Gastrointestina tract	Dental plaque; Periodontal infection
Prochlorococcus marinus AS9601	non host-associated	no		Arabian Sea at a depth of 50m on November 1995		None
Prochlorococcus marinus MIT9211	non host-associated	no		Equatorial Pacific at a depth of 83m on April 1992		None
Prochlorococcus marinus MIT9215	non host-associated	no		5m depth at Equatorial Pacific on October 1992		None
Prochlorococcus marinus MIT9301	non host-associated	no		Sargasso Sea at a depth of 90m on July 1993		None
Prochlorococcus marinus MIT9303	non host-associated	no		Sargasso Sea at a depth of 100m on July 1992		None
Prochlorococcus marinus MIT9312	non host-associated	no		Gulf Stream of North Atlantic Ocean		None
Prochlorococcus marinus MIT9515	non host-associated	no		Equatorial Pacific on June 1995; Surface waters		None

Organism*	Animal host-associated	Mucosa associated**	M60-like domain	Isolation site	Body sample site	Disease
Prochlorococcus marinus NATL1A	non host-associated	no		North Atlantic Ocean at a depth of 30m on April 1990		None
Prochlorococcus marinus NATL2A	non host-associated	no		30 meters depth in North Atlantic Ocean		None
Prochlorococcus marinus pastoris CCMP1986 (MED4)	non host-associated	no		5m depth in Mediterranean Sea		None
Prochlorococcus sp. CC9311	non host-associated	no		Oligotrophic edge of the California Current at 95m depth in 1993; CalCOFI cruise 93204; station 83.110		None
Prochlorococcus sp. CC9605 (oligotrophic)	non host-associated	no		51m depth of California coast on 1996		None
Prochlorococcus sp. CC9902 (coastal)	non host-associated	no		5m depth in California current on 1999		None
Prochlorococcus sp. WH 7803	non host-associated	no		Sargasso Sea;North Atlantic		None
Prochlorococcus sp. WH8102	non host-associated	no		Sargasso Sea from Oceanus cruise 92; in 1981		None
Propionibacterium acnes KPA171202	host-associated			Human skin	Skin	Acne
Proteus mirabilis HI4320	host-associated	yes		Urine of a nursing home patient with a long term (>30 day) indwelling urinary catheter	Urogenital tract	Urinary tract infection; Urolithiasis; Ulcer; Encephalitis; Pneumonia; Pyelonephritis; Septicemia; Surgical wound infection
Pseudoalteromonas atlantica T6c	host-associated			Lesions on crabs with shell disease		None
Pseudoalteromonas haloplanktis TAC125	non host-associated	no		Coastal sea water near a French Antarctic station; Adelia Land		None
Pseudomonas aeruginosa LESB58	host-associated	yes	+	Liverpool Cystic Fibrosis (CF) clinic center	Airways	Cystic Fibrosis
Pseudomonas aeruginosa PA14 UCBPP	host-associated		+	Human clinical isolate		Nocosomial infection; Opportunistic infection
Pseudomonas aeruginosa PA7	host-associated		+	clinical isolate		Opportunistic infection
Pseudomonas aeruginosa PAO1	host-associated		+			Nosocomial infection
Pseudomonas entomophila L48	host-associated					Cellular destruction
Pseudomonas fluorescens Pf0-1	non host-associated			Agricultural loam soil in 1988		None
Pseudomonas mendocina ymp	host-associated			Argentina	Blood	Endocarditis; Spondylodiscitis
Pseudomonas putida F1	non host-associated			Polluted creek in Urbana IL		None
Pseudomonas putida KT2440	non host-associated			Derived from a toluene-degrading isolate		None
Pseudomonas putida W619	non host-associated	no		Black Cottonwood tree		None
Pseudomonas stutzeri A1501	non host-associated	no		Rice roots that had been inoculated with strain A15 in a rice paddy in China		None
Pseudomonas syringae phaseolicola 1448A / Race 6	non host-associated					Halo blight; Plant rot
Pseudomonas syringae syringae B728a	non host-associated		+			Plant rot
Pseudomonas syringae tomato DC3000	non host-associated	no	+	tomato; Channel Islands; Guernsey; UK		Plant rot; Speck disease

Organism*	Animal host-associated	Mucosa associated**	M60-like domain	Isolation site	Body sample site	Disease
Psychrobacter arcticum 273-4	non host-associated	no		20-40000 year old Serbian permafrost in Russia; Siberia; Kolyma lowland		None
Psychrobacter cryohalolentis K5	non host-associated	no		Saline liquid found 11-24 m below the surface within a forty thousand-year-old Siberian permafrost at the Kolyma-Indigirka lowland in Siberia		None
Psychromonas ingrahamii 37	non host-associated	no		Sea ice off Point Barrow in northern Alaska		None
Pyrobaculum aerophilum IM2	non host-associated	no		Boiling marine water hole in Maronti Beach; Ischia; Italy		None
Pyrobaculum arsenaticum PZ6	non host-associated	no		Hot spring at Pisciarelli Solfatara Naples Italy		None
Pyrobaculum calidifontis JCM 11548	non host-associated	no		Terrestrial hot spring in the Philippines		None
Pyrobaculum islandicum DSM 4184	non host-associated	no		Water from a geothermal power plant in Iceland		None
Pyrococcus abyssi GE5	non host-associated	no		Active chimney in the North Fiji Basin of the Pacific Ocean at a depth of 3500m		None
Pyrococcus furiosus JCM 8422	non host-associated	no		Shallow marine solfataric region at Vulcano Island Italy		None
Pyrococcus horikoshii OT3	non host-associated	no		Hydrothermal vent at Okinawa Trough in the Pacific Ocean at a depth of 1395m		None
Ralstonia eutropha H16	non host-associated	no		Sludge		None
Ralstonia pickettii 12J	host-associated	no		copper-contaminated sediment from a lake in Michigan		Nosocomial infection
Ralstonia solanacearum GMI1000	non host-associated	no		wilted tomato plant; French Guyana		Plant rot; Wilting disease
Renibacterium salmoninarum ATCC 33209	host-associated			Yearling chinook salmon at a salmon hatchery in Western Oregon		Bacterial kidney disease
Rhizobium etli CFN42	non host-associated			Phaseolus vulgaris; Guanajuato Mexico		None
Rhizobium leguminosarum WSM2304	non host-associated			Glencoe Research Station; INIA Uruguay		None
Rhodobacter sphaeroides KD131	non host-associated	no		Sea mud off the coast of DaeBu Island; South Korea		
Rhodococcus sp RHA1	non host-associated	no		Soil contaminated with gamma-hexachlorocyclohexane in Japan		None
Rhodoferax ferrireducens T118	non host-associated	no		Aquifer sediment collected at a depth of 18 feet		None
Rhodopirellula baltica SH 1	non host-associated			Kieler Bucht (a fiord near the city of Kiel in Germany)		None
Rhodopseudomonas palustris BisA53	non host-associated	no		freshwater sediment samples from De Biesbosch and Haren; the Netherlands		None
Rhodopseudomonas palustris BisB18	non host-associated	no		freshwater sediment samples from De Biesbosch and Haren; the Netherlands		None
Rhodopseudomonas palustris BisB5	non host-associated	no		freshwater sediment samples from De Biesbosch and Haren; the Netherlands		None

Organism*	Animal host-associated	Mucosa associated**	M60-like domain	Isolation site	Body sample site	Disease
Rhodopseudomonas palustris HaA2	non host-associated			Haren		None
Rhodopseudomonas palustris TIE-1	non host-associated	no		Iron-rich mat from School Street Marsh in Woods Hole MA		None
Rhodospirillum centenum SW	non host-associated			derived from ATCC 43720		None
Rickettsia akari Hartford	host-associated					Rickettsial pox
Rickettsia bellii OSU 85-389	host-associated			Isolated in Vero cell culture at 34 C by Karl Poetter and Chip Pretzman; spaghetti forms seen		Epidemic typhus
Rickettsia bellii RML369-C	host-associated			Embryonated chicken eggs from a triturated pool of unfed adult Dermacentor variabilis ticks collected from vegetation near Fayetteville Arkansas USA in 1966		None
Rickettsia canadensis McKiel	host-associated	no		Ticks in Canada		Epidemic typhus
Rickettsia conorii Malish 7	host-associated			Human in South Africa		Rocky Mountain Spotted Fever
Rickettsia felis URRWX-Cal2	host-associated					Flea-borne Spotted Fever; Rickettsiosis
Rickettsia massiliae MTU5	host-associated			Rhipicephalus turanicus ticks collected on horses in Camargues; France		None
Rickettsia prowazekii Madrid E	host-associated			Typhus patient in Madrid		Typhus; Rocky Mountain Spotted Fever
Rickettsia rickettsii Iowa	host-associated					Rocky Mountain Spotted Fever
Rickettsia rickettsii Sheila Smith	host-associated			Patient with Rocky Mountain spotted fever		Rocky Mountain Spotted Fever
Rickettsia typhi Wilmington	host-associated			blood from patient; North Carolina; 1928	Blood	Typhus
Roseiflexus castenholzii HLO8; DSM 13941	non host-associated	no		Hot spring microbial mat		None
Roseiflexus sp RS-1	non host-associated	no		Hot spring microbial mat		None
Roseobacter denitrificans OCh 114	non host-associated	no		seaweed; Enteromorpha linza from Aburatsubo Inlet Kanagawa Japan		None
Rubrobacter xylanophilus DSM 9941	non host-associated	no		Thermally polluted industrial runoff in the United Kingdom		None
Saccharomyces cerevisiae	non host-associated			Baker strain		
Saccharophagus degradans 2-40	non host-associated	no		Decaying Spartina alterniflora a salt marsh cord grass in the Chesapeake Bay		None
Salinibacter ruber M31	non host-associated	no		Saltern crystallizer ponds in Spain		None
Salinispora arenicola CNS205	non host-associated	no		Beach sand at a depth of 1 meter from Sweetings Cay in the Bahamas		None
Salinispora tropica CNB-440	non host-associated	no		Coarse beach sand off the Bahamas		None
Salmonella enterica Agona SL483	host-associated	yes				Salmonellosis; Food poisoning; Gastroenteritis
Salmonella enterica arizonae sv 62:z4;z23 RSK2980	host-associated	yes	+	cornsnake in 1986 in Oregon		Food poisoning; Salmonellosis; Gastroenteritis

Organism*	Animal host-associated	Mucosa associated**	M60-like domain	Isolation site	Body sample site	Disease
Salmonella enterica Choleraesuis SC-B67	host-associated	yes		58-year old man with sepsis		Swine paratyphoid; Food poisoning; Gastroenteritis; Salmonellosis
Salmonella enterica enterica PT4 P125109	host-associated	yes	+	Outbreak of human food-poisoning in the UK which was traced back to a poultry farm.		Salmonellosis; Food poisoning; Gastroenteritis
Salmonella enterica Gallinarum 287/91	host-associated	yes		outbreak of fowl typhoid in brownegg laying hens by Prof. A. Berchieri; University of Sao Paulo; Jaboticabal; Brazil		Salmonellosis; Food poisoning; Gastroenteritis
Salmonella enterica Newport SL254	host-associated	yes		MDR strain from one of two distinct lineages of the Newport serovar		Salmonellosis; Food poisoning; Gastroenteritis
Salmonella enterica Schwarzengrund CVM19633	host-associated	yes	+			Salmonellosis; Food poisoning; Gastroenteritis
Salmonella enterica sv Dublin CT_02021853	host-associated	yes	+	bovine-adapted serovar		Salmonellosis; Food poisoning; Gastroenteritis
Salmonella enterica sv Heidelberg SL476; CVM30485	host-associated					Salmonellosis; Food poisoning; Gastroenteritis
Salmonella enterica sv Paratyphi A AKU_12601	host-associated	yes		clinical isolate from a child with paratyphoid fever in Karachi; Pakistan; 2004		Typhoid fever; Food poisoning; Gastroenteritis; Salmonellosis
Salmonella enterica sv Paratyphi A SARB42	host-associated	yes				Typhoid fever; Food poisoning; Gastroenteritis; Salmonellosis
Salmonella enterica sv Paratyphi B SPB7	host-associated	yes		Patient in Malaysia in 2002		Salmonellosis; Food poisoning; Paratyphoid fever
Salmonella enterica sv Typhi CT18	host-associated	yes				Typhoid fever; Food poisoning; Salmonellosis
Salmonella enterica Typhi Ty2 ATCC700931	host-associated	yes				Typhoid fever; Food poisoning; Gastroenteritis; Salmonellosis
Salmonella enterica Typhimurium LT2 SGSC1412 / LT2	host-associated	yes		1940s by Lilleengen		Salmonellosis; Food poisoning; Gastroenteritis
Serratia proteamaculans 568	host-associated	yes				Pneumonia
Shewanella amazonensis SB2B	non host-associated	no		Shallow marine deposits of the Amazon River delta off of the coast of Brazil		None
Shewanella baltica OS155	non host-associated	no	+	Sea-water; oxic zone; 2 ml per litre of oxygen; 90m depth from Baltic Sea		None
Shewanella baltica OS185	non host-associated	no	+	Sea-water; anoxic interface; 120 m depth from the Baltic Sea		None
Shewanella baltica OS195	non host-associated	no	+	Sea-water; anoxic zone; 140 m depth from the Baltic Sea		None
Shewanella baltica OS223	non host-associated	no	+	Sea water; oxic-anoxic interface; 120m depth from the Baltic Sea		None
Shewanella denitrificans OS217	non host-associated	no		Gotland Deep an anoxic basin in the central Baltic Sea in 1986 from a depth of 120-130m		None
Shewanella frigidimarina NCIMB 400	non host-associated	no		North Sea near Aberdeen United Kingdom		None

Organism*	**Animal host-associated**	**Mucosa associated****	**M60-like domain**	**Isolation site**	**Body sample site**	**Disease**
Shewanella halifaxensis HAW-EB4	non host-associated	no	+	Sediment of the Emerald Basin at 215m depth offshore of Halifax Harbour in the Atlantic Ocean		None
Shewanella loihica PV-4	non host-associated	no	+	Iron-rich mat; hydrothermal vent; 1;325 m depth from Naha Vents; on the south rift of Loihi; Hawaii		None
Shewanella oneidensis MR-1	non host-associated	no	+	Sediment; anaerobic; Mn(IV) reduction; Oneida lake in New York		None
Shewanella pealeana ANG-SQ1; ATCC 700345	host-associated	yes	+	Microbial community colonizing the accessory nidamental gland of the squid Loligo pealei; from Woods Hole Harbor; Massachusetts		None
Shewanella putrefaciens CN-32	non host-associated	no		Subsurface; shale sandstone; 250 m depth from Albuquerque; New Mexico		None
Shewanella sediminis HAW-EB3	non host-associated	no		Sediment at depth of 215m from an unexploded-ordinance-dumping site at Halifax		None
Shewanella sp. ANA-3	non host-associated	no	+	Arsenate treated wood pier that was in a brackish estuary (Eel Pond) in Woods Hole Massachusetts		None
Shewanella sp. MR-4	non host-associated	no	+	Sea-water; oxic zone; 16oC; 5 m depth in the Black sea		None
Shewanella sp. MR-7	non host-associated	no	+	Sea-water; anoxic zone; high NO3; 60 m depth in the Black sea		None
Shewanella sp. W3-18-1	non host-associated	no		Marine sediment; under 997 m depth of oxic water from the Washington coast; Pacific Ocean		None
Shewanella woodyi MS32; ATCC 51908	non host-associated	no	+	Sediment; 5;110 m depth from the Strait of Gibraltar; Mediterranean Sea		None
Shigella boydii BS512;CDC 3083-94	host-associated	yes		12-year-old boy in Arizona by Dr. Nancy Stockbine		Dysenteria; Food poisoning
Shigella boydii Sb227	host-associated	yes		Epidemic in China in 1950s		Dysenteria; Food poisoning
Shigella dysenteriae Sd197	host-associated	yes		Epidemic in China in 1950s		Dysenteria; Food poisoning
Shigella flexneri 2457T	host-associated	yes				Gastroenteritis; Dysenteria; Food poisoning
Shigella flexneri 301	host-associated	yes		In 1984 from a patient in Beijing China		Shigellosis; Dysenteria; Food poisoning
Shigella flexneri 5b 8401	host-associated	yes				Shigellosis; Dysenteria; Food poisoning
Shigella sonnei Ss046	host-associated	yes		Epidemic in China in 1950s		Dysenteria; Food poisoning
Silicibacter sp. TM1040	non host-associated			culture of the dinoflagellate Pfiesteria piscicida CCMP1830		None
Sinorhizobium medicae WSM419	non host-associated			Forestry Station 7 k south of Tempio; Sardinia		None
Sinorhizobium meliloti 1021	non host-associated			Streptomycin resistant derivative of strain 2011		None

Organism*	Animal host-associated	Mucosa associated**	M60-like domain	Isolation site	Body sample site	Disease
Solibacter usitatus Ellin6076	non host-associated	no		rotationally grazed pasture of perennial ryegrass and white clover in Victoria; Australia		None
Sphingomonas wittichii RW1	non host-associated	no		Elbe River in Germany		None
Sphingopyxis alaskensis RB2256	non host-associated			Resurrection Bay in the Gulf of Alaska		None
Staphylococcus aureus aureus COL	host-associated					Toxic-shock syndrome; Styes; Pneumonia; Phlebitis; Osteomyelitis; Nosocomial infection; Boils; Mastitis; Impetigo; Furunculosis; Fever; Endocarditis; Meningitis
Staphylococcus aureus aureus JH9 VISA	host-associated			Patient undergoing vancomycin treatment		Toxic-shock syndrome; Staphylococcal scarlet fever
Staphylococcus aureus aureus MRSA USA300; FPR3757	host-associated	yes				Pneumonia; Septicemia
Staphylococcus aureus aureus MRSA252	host-associated					Staphylococcal scarlet fever; Toxic-shock syndrome
Staphylococcus aureus aureus MSSA476	host-associated					Staphylococcal scarlet fever; Toxic-shock syndrome
Staphylococcus aureus aureus Mu3	host-associated	yes		sputum from a lung cancer patient with MRSA pneumonia; Japan	Airways	Mastitis; Nosocomial infection
Staphylococcus aureus aureus Mu50 (VRSA)	host-associated	yes		Pus of a Japanese male baby with a surgical wound infection that did not respond to vancomycin in 1997		Pneumonia; Phlebitis; Osteomyelitis; Deep abscesses; Meningitis; Mastitis; Endocarditis; Nosocomial infection
Staphylococcus aureus aureus MW2	host-associated			Sour cassava in Nigeria		Staphylococcal scarlet fever; Toxic-shock syndrome
Staphylococcus aureus aureus N315 (MRSA)	host-associated	yes		Pharyngeal smear of a Japanese patient in 1982	Airways	Pneumonia; Phlebitis; Osteomyelitis; Deep abscesses; Meningitis; Mastitis; Endocarditis; Nosocomial infection
Staphylococcus aureus aureus NCTC 8325	host-associated					Staphylococcal scarlet fever; Toxic-shock syndrome
Staphylococcus aureus aureus Newman	host-associated					Skin infection; Endocarditis; Pneumonia
Staphylococcus aureus aureus USA300_TCH1516	host-associated	yes		Human skin	Skin	Pneumonia; Septicemia
Staphylococcus aureus JH1 VISA	host-associated					Staphylococcal scarlet fever; Toxic-shock syndrome
Staphylococcus aureus RF122 bovine	host-associated					Mastitis; Nosocomial infection

Organism*	Animal host-associated	Mucosa associated**	M60-like domain	Isolation site	Body sample site	Disease
Staphylococcus epidermidis RP62A	host-associated			Patient with intravascular catheter-associated sepsis		Toxic-shock syndrome; Nosocomial infection; Septicemia; Staphylococcal scarlet fever
Staphylococcus haemolyticus JCSC1435	host-associated			Japanese inpatient at Juntendo Hospital; Tokyo; in 2000		Opportunistic infection
Staphylococcus saprophyticus saprophyticus GTC 265	host-associated	yes		Human urine specimen	Urogenital tract	Urinary infection
Staphylothermus marinus F1; DSM 3639	non host-associated	no		Hydrothermal marine sediment from Vulcano Island in Italy		None
Stenotrophomonas maltophilia K279a	host-associated			Blood of a elderly male patient undergoing chemotherapy at the Bristol Oncology Unit; Bristol; UK in 1998	Blood	Pulmonary infection; Bacteremia; Nosocomial infection; Opportunistic infection
Stenotrophomonas maltophilia R551-3	host-associated	yes				Urinary infection; Respiratory infection; Nosocomial infection; Blood infection
Streptococcus agalactiae 2603V/R	host-associated	yes		clinical isolate		Septicemia; Meningitis; Pneumonia
Streptococcus agalactiae A909;ATCC BAA-1138	host-associated	yes				Septicemia; Meningitis; Pneumonia
Streptococcus agalactiae NEM316	host-associated			Case of fatal septicemia	Blood	Meningitis
Streptococcus equi zooepidemicus MGCS10565	host-associated	yes		throat of a patient with nephritis iagnosed during an epidemic in the state of Minas Gerais; Brazil	Airways	Opportunistic infection
Streptococcus gordonii Challis CH1	host-associated	yes				Periodontal infection; Dental plaque; Endocarditis
Streptococcus mutans UA159	host-associated	yes		Child with active dental caries in 1982	Oral	Dental caries
Streptococcus pneumoniae 23F ST81; ATCC 700669	host-associated	yes				Pneumonia
Streptococcus pneumoniae CGSP14	host-associated	yes				Pneumonia
Streptococcus pneumoniae D39	host-associated	yes				Pneumonia; Meningitis; Otitis media
Streptococcus pneumoniae G54 (MLST ST63)	host-associated	yes		Genova Italy by G. Schito from a respiratory sample in 1997	Airways	Pneumonia
Streptococcus pneumoniae Hungary 19A-6	host-associated	yes		Human ear; Hungary	Ear	Pneumonia
Streptococcus pneumoniae R6	host-associated	yes				Pneumonia
Streptococcus pneumoniae TIGR4	host-associated	yes		Blood of a 30 year old male patient in Kongsvinger Norway	Blood	Pneumonia; Meningitis; Otitis media
Streptococcus pyogenes M18 MGAS8232	host-associated					Rheumatic fever
Streptococcus pyogenes M3 (SSI-1)	host-associated			Toxic-shock patient in Japan		Necrotizing fasciitis; Rheumatic fever
Streptococcus pyogenes M49; NZ131	host-associated			Patient with acute glomerulonephritis and was provided by Diana Martin; New Zealand Communicable Diseases Center; Porirua; New Zealand		Glomerulonephritis

Organism*	**Animal host-associated**	**Mucosa associated****	**M60-like domain**	**Isolation site**	**Body sample site**	**Disease**
Streptococcus pyogenes M6 MGAS10394	host-associated	yes				Tonsillitis; Pharyngitis; Rheumatic fever
Streptococcus pyogenes Manfredo (M5)	host-associated			Patient in the 1950's in Chicago		Rheumatic fever
Streptococcus pyogenes MGAS2096	host-associated					Rheumatic fever
Streptococcus pyogenes MGAS315	host-associated					Rheumatic fever
Streptococcus sanguinis SK36	host-associated					Endocarditis
Streptococcus suis 05ZYH33	host-associated			Chinese virulent strain isolated from fatal cases of STSS in 2005		Septicemia; Arthritis; Endocarditis; Meningitis
Streptococcus suis 98HAH33	host-associated			Chinese virulent strain isolated from fatal cases of STSS in 1998		Septicemia; Arthritis; Endocarditis; Meningitis
Streptococcus thermophilus LMG18311	non host-associated	no		Commercial yogurt in 1974 in the United Kingdom		None
Streptococcus uberis 0140J	host-associated			clinical bovine mastitis case		Mastitis
Streptomyces avermitilis MA-4680	non host-associated	no		Soil sample collected in Shizuoka Prefecture Japan		None
Streptomyces coelicolor A3(2) M145	non host-associated	no		Derivative of the laboratory strain A3(2)		None
Sulfolobus acidocaldarius DSM 639	non host-associated	no		Terrestrial solfataras		None
Sulfolobus tokodaii 7; JCM 10545	non host-associated	no		Beppu Hot Springs in the geothermal area of Kyushu Island Japan		None
Sulfurihydrogenibium sp. YO3AOP1	non host-associated			Soil from Obsidian Pool in Yellowstone National Park; USA		None
Sulfurimonas denitrificans ATCC 33889	non host-associated	no		Estuarine mud in Netherlands		None
Sulfurovum sp NBC37-1	non host-associated	no		30-m-tall sulfide mound in the Iheya North field; Japan (water depth; 1;000 m)		None
Symbiobacterium thermophilum IAM 14863	non host-associated			Compost in Hiroshima; Japan		None
Synechococcus sp JA-2-3B'a(2-13)	non host-associated	no		Top 2mm of microbial mat samples from Octopus Spring Yellowstone National Park		None
Synechococcus sp. PCC 7002	non host-associated	no		1961 from a mud sample that came from the fish pens from Magueyes Island in Puerto Rico		None
Synechococcus sp. RCC307	non host-associated	no		Seawater taken at a depth of 15 meters from the Mediterranean Sea		None
Synechocystis sp. PCC6803	non host-associated	no		Freshwater lake in 1968		None
Syntrophobacter fumaroxidans MPOB	non host-associated	no		Granular sludge from an anaerobic sludge bed reactor; The Netherlands		None
Syntrophus aciditrophicus SB	non host-associated	no		Sludge from a sewage treatment plant in Norman Oklahoma		None
Thauera sp. MZ1T	non host-associated	no		Wastewater treatment plant		None
Theileria parva Muguga	host-associated					East Coast Fever

Organism*	Animal host-associated	Mucosa associated**	M60-like domain	Isolation site	Body sample site	Disease
Thermoanaerobacter ethanolicus X514	non host-associated	no		Anaerobic enrichment culture from a deep subsurface sample (2000 m below the surface) taken from a core hole at the Piceance Basin; Colorado; USA		None
Thermoanaerobacter pseudoethanolicus 39E	non host-associated	no		Thermal springs in Yellowstone National Park		None
Thermoanaerobacter tengcongensis MB4T / JCM 11007	non host-associated	no		Hot spring in Tengcong China		None
Thermobifida fusca YX	host-associated			compost pile in the 1970's		Respiratory infection; Farmer's lung; Mushroom worker's disease
Thermococcus kodakaraensis KOD1	non host-associated	no		Solfatara on Kodakara Island Kagoshima; Japan		None
Thermococcus onnurineus NA1	non host-associated	no		PACMANUS hydrothermal vent sediment at a depth of 1650 meters		None
Thermodesulfovibrio yellowstonii DSM 11347	non host-associated	no		Thermal vent in Yellowstone Lake in Wyoming		None
Thermofilum pendens Hrk 5	non host-associated	no		Solfataric hot spring in Iceland		None
Thermomicrobium roseum DSM 5159	non host-associated	no		Toadstool Spring in Yellowstone National Park		None
Thermoplasma acidophilum DSM 1728	non host-associated	no		Self-heating coal refuse pile in southwestern Indiana		None
Thermoplasma volcanium GSS1	non host-associated	no		Acidic hydrothermal vents on the shore of Aeolian Island of Vulcano Italy		None
Thermoproteus neutrophilus V24Sta	non host-associated	no		Hot spring in Iceland		None
Thermosipho africanus TCF52B	non host-associated	no		Troll oil formation in the North Sea		None
Thermosipho melanesiensis BI429	non host-associated	no		Gills of the deep-sea vent hydrothermal mussel Bathymodiolus brevior from the Lau Basin at Southwestern Pacific Ocean; between 1832 and 1887 metres (latitude; 22o329S; longitude; 176o439W)		None
Thermosynechococcus elongatus BP-1	non host-associated	no		Beppu hot spring in Japan		None
Thermotoga lettingae TMOT	non host-associated	no		Methanol-degrading; sulfate-reducing bioreactor		None
Thermotoga maritima MSB8	non host-associated	no		Geothermal marine area near Vulcano Italy		None
Thermotoga neapolitana DSM 4359	non host-associated	no		Black smoker in the bay near Naples Italy in 1986		None
Thermotoga petrophila RKU-1	non host-associated			Production waters of the Kubiki oil reservoir in Niigata Japan		None
Thermotoga sp. RQ2	non host-associated	no		Geothermally heated seafloor in the Azores		None
Thermus thermophilus HB27	non host-associated	no		Thermal vent in Japan		None
Thermus thermophilus HB8	non host-associated	no		Thermal vent in Japan		None
Thioalkalivibrio sp HL-EbGR7	non host-associated	no		Sulfide-oxidizing bioreactor		None
Thiobacillus denitrificans ATCC 25259	non host-associated	no		Soil from Texas USA		None

Organism*	**Animal host-associated**	**Mucosa associated****	**M60-like domain**	**Isolation site**	**Body sample site**	**Disease**
Thiomicrospira crunogena XCL-2	non host-associated	no		Deep-sea hydrothermal vent		None
Treponema denticola ATTC35405	host-associated	yes				Periodontal infection
Treponema pallidum pallidum Nichols	host-associated			Neurosyphilitic patient in 1912		Syphilis
Treponema pallidum pallidum SS14	host-associated	yes		Patient in Atlanta with secondary syphilis who did not respond to erythromycin therapy		Syphilis
Trichodesmium erythraeum IMS101	non host-associated	no		coastal waters; North Carolina USA		None
Trichomonas vaginalis G3	host-associated	yes	+			Sexually transmitted disease; Trichomoniasis
Tropheryma whipplei TW08/27	host-associated			Cerebrospinal fluid a woman who had suffered severe weight loss in Germany	Brain	Whipple's disease
Tropheryma whipplei Twist	host-associated					Whipple's disease
Trypanosoma brucei TREU927/4 GUTat10.1	host-associated	no				Human sleeping sickness; Trypanosomiasis
Ureaplasma parvum sv 3; ATCC 27815	host-associated	yes				Respiratory infection; Urinary tract infection
Ureaplasma urealyticum (parvum) sv 3; ATCC 700970	host-associated	yes			Urogenital tract	Supperative arthritis; Sexually transmitted disease; Meningitis; Pneumonia; Septicemia
Ureaplasma urealyticum Western; sv 10; ATCC 33699	host-associated	yes			Urogenital tract	Non-specific urethritis (NSU); Infertility; Chorioamnioitis
Ustilago maydis 521	non host-associated					Corn smut
Ustilago maydis FB1	non host-associated					Corn smut
Verminephrobacter eiseniae EF01-2	non host-associated	no		Kidney of the earthworm Eisenia foetida		None
Vibrio cholerae N16961; Biotype ElTor	host-associated	yes	+	stool from cholera patient in Epidemic outbreak in Bangladesh in 1971	Gastrointestina tract	Food poisoning; Cholera; Diarrhea
Vibrio cholerae O395	host-associated	yes	+	6th pandemic isolate		Cholera; Food poisoning
Vibrio fischeri ES114	host-associated		+	Light organs of the squid Euprymna scolopes		None
Vibrio fischeri MJ11	host-associated		+	Squid light organ in Japan		None
Vibrio harveyi BB120; ATCC BAA-1116	non host-associated			Ocean isolate obtained in 1993		None
Vibrio parahaemolyticus RIMD 2210633	host-associated	yes	+	Clinical strain isolated in 1996 in Osaka Japan		Gastroenteritis
Vibrio splendidus LGP32	host-associated	yes	+			Vibriosis
Vibrio vulnificus CMCP6	host-associated	yes	+			Gastroenteritis; Septicemia
Vibrio vulnificus YJ016	host-associated	yes	+	Hospital isolate from Taiwan		Septicemia; Gastroenteritis; Food poisoning
Wolbachia pipientis (Culex quinquefasciatus) Pel wPip	non host-associated			preblastoderm embryos of the Pel strain of Culex pipiens mosquitoes		None
Wolinella succinogenes DSM 1740	host-associated	yes		Bovine rumen fluid	Gastrointestina tract	None
Xanthomonas axonopodis pv. citri XV101; 306	non host-associated					Citrus canker

Organism*	**Animal host-associated**	**Mucosa associated****	**M60-like domain**	**Isolation site**	**Body sample site**	**Disease**
Xanthomonas campestris campestris 8004	non host-associated	no		Inflected cauliflower in Sussex UK in 1958		Black rot; Citrus canker
Xanthomonas campestris campestris ATCC 33913	non host-associated	no		Cabbage		Black rot
Xanthomonas campestris campestris B100	non host-associated	no				Black rot
Xanthomonas campestris vesicatoria 85-10	non host-associated	no				Bacterial spot
Xanthomonas oryzae MAFF 311018	non host-associated	no				Leaf blight; Rice blight
Xanthomonas oryzae pv. oryzae KACC10331	non host-associated	no				Blight disease
Xanthomonas oryzae pv. oryzae PXO99A	non host-associated	no		5-azacytidine resistant derivative of PXO99; isolated in Los Baqos and classified as Philippine race 6		Rice blight
Xylella fastidiosa CVC 8.1.b clone 9.a.5.c	non host-associated	no		Infected twigs derived from the sweet orange strain Valencia in Brazil		Citrus variegated chlorosis; Pierces disease
Xylella fastidiosa M12	non host-associated					Citrus variegated chlorosis
Xylella fastidiosa M23	non host-associated	no		Almond tree in California		Citrus variegated chlorosis
Xylella fastidiosa-grape Temecula1	non host-associated	no		In 1998 from a naturally infected Californian grapevine		Black rot; Citrus canker
Yersinia enterocolitica 8081	host-associated	yes	+			Gastroenteritis; Food poisoning
Yersinia pestis Angola	host-associated	yes	+			Gastroenteritis; Bubonic plague; Food poisoning
Yersinia pestis Antiqua	host-associated		+	Soil sample from the Republic of Congo		Bubonic and Pneumonic plague
Yersinia pestis Mediaevalis KIM10+	host-associated					Bubonic plague
Yersinia pestis Microtus 91001	host-associated		+			Bubonic plague
Yersinia pestis Nepal516	host-associated			Soil sample from Nepal		Bubonic and Pneumonic plague
Yersinia pestis Orientalis CO-92	host-associated		+	Patientin the USA who died of pneumonic plague after acquiring the disease from an infected cat	Airways	Bubonic plague
Yersinia pestis Pestoides F	host-associated		+			Bubonic and Pneumonic plague
Yersinia pseudotuberculosis IP 31758	host-associated	yes	+	Stools of a patient presenting with scarlet-like fever in the Primorski region of the former USSR on 1996	Gastrointestina tract	Food poisoning; Gastroenteritis
Yersinia pseudotuberculosis IP32953	host-associated	yes	+			Food poisoning; Gastroenteritis
Yersinia pseudotuberculosis PB1/+	host-associated	yes	+			Gastroenteritis

Appendix J

Summary of proteins containing M60-like domains, their taxa distribution and sequence features

The presence of the HEXXH zincin motif, alpha-helix transmembrane (TM) region and N-terminal SPI or SPII cleavage sites for the M60-like-possessing proteins from each taxa are indicated. Proteins containing M60-like domains were identified by using HMMER search with the PF13402 profile at cut-off e-value of less than 1×10^{-5}.

Taxa possessing the M60-like domain	**Habitat or isolation source**	**Disease via or in**	**Number of domain per strain (total per species)**	**HEXXH motif**	**TM**	**SPase**
Bacteria						
Firmicutes						
*Bacillus anthracis**	Soil/IG	GIT, RT	1-2(30)	Yes	-	Yes
Bacillus cereus	Soil/GIT/IG	GIT, RT	1-5(91)	+(90)	+(30)	+(80)
*Bacillus mycoides**	Soil, other (animal,plants)????	Plant disease	1-2(3)	Yes	-	+(2)
Bacillus pseudomycoides DSM 12442	Soil	-	2	Yes	+(1)	+(1)
*Bacillus thuringiensis**	Soil/IG	IG	1-4(47)	Yes	-	+(36)
Bacillus weihenstephanensis KBAB4	Soil	-	3	Yes	-	Yes
Clostridium bartlettii DSM 16795	GIT	-	2	Yes	-	Yes
*Clostridium botulinum**	GIT	GIT	1(7)	Yes	-	Yes
Clostridium difficile QCD-32g58			1	Yes	-	-
Clostridium hathewayi DSM 13479	GIT	-	1	Yes	-	Yes
Clostridium hiranonis DSM 13275	GIT	-	1	Yes	-	Yes
*Clostridium perfringens**	Soil/GIT	GIT	1-6(18)	+(16)	-	+(16)
Clostridium ramosum DSM 1402	GIT	Opportunistic infections	1	Yes	-	Yes
Clostridium sp. 7_2_43FAA			1	Yes	-	Yes

Taxa possessing the M60-like domain	**Habitat or isolation source**	**Disease via or in**	**Number of domain per strain (total per species)**	**HEXXH motif**	**TM**	**SPase**
Eubacterium dolichum DSM 3991	GIT	-	1	Yes	-	Yes
Geobacillus sp. Y412MC10	Soil	-	2	Yes	-	Yes
*Lactobacillus jensenii**	UGT	-	1(3)	Yes	-	+(2)
Listeria grayi DSM 20601	GIT	-	1	Yes	-	Yes
Mollicutes bacterium D7	GIT		1	Yes	-	Yes
Paenibacillus larvae subsp. larvae BRL-230010		IG	3	Yes	-	+(1)
Subdoligranulum variabile DSM 15176	GIT	-	1	Yes	-	Yes
Tenericutes						
Mycoplasma penetrans HF-2		UGT, RT	3	+(2)	Yes	-
Bacteroidetes						
Bacteroides caccae ATCC 43185	GIT	GIT	16	Yes	+(1)	+(14)
Bacteroides coprophilus DSM 18228	GIT	-	1	Yes	-	Yes
Bacteroides finegoldii DSM 17565			1	-	-	Yes
*Bacteroides fragilis**	GIT	GIT	1(3)	Yes	-	Yes
Bacteroides plebeius DSM 17135	GIT	-	2	Yes	-	Yes
*Bacteroides sp.**	GIT	-	1-4(7)	+(5)	+(1)	+(5)
Bacteroides thetaiotaomicron VPI-5482	GIT	GIT	4	+(3)	+(1)	+(3)
Bacteroides vulgatus ATCC 8482	GIT	GIT	1	Yes	-	Yes
Chitinophaga pinensis DSM 2588	Soil - pine litter	-	1	Yes	-	Yes
Prevotella melaninogenica ATCC 25845	Oral cavity	Periodontal disease	1	Yes	-	Yes
Sphingobacterium spiritivorum	Water	RT	6(12)	Yes	-	Yes
Actinobacteria						
Brachybacterium faecium DSM 4810	Poultry deep litter		1	Yes	-	Yes
Eggerthella lenta DSM 2243	GIT	Bacteremia (rare)	1	Yes	-	Yes
Proteobacteria						
Escherichia albertii TW07627			1	Yes	-	Yes
*Escherichia coli**	GIT	GIT, UGT	1(27)	Yes	-	+(22)
Escherichia fergusonii ATCC 35469	GIT	UGT, wound	1	Yes	-	Yes
*Escherichia sp.**	GIT	-	1(3)	Yes	-	+(2)
Grimontia hollisae CIP 101886	-	GIT	2	Yes	-	-
Hahella chejuensis KCTC 2396	Marine sediment	-	1	-	-	Yes
Pantoea sp. At-9b			1	Yes	-	Yes
Photobacterium damselae subsp. damsela	Marine	Skin of fish and human	1	Yes	-	Yes
Photorhabdus asymbiotica	Entomopathogenic nematode Heterorhabditis indica		1	Yes	-	-
*Pseudomonas aeruginosa**			7	Yes	-	Yes
Pseudomonas syringae	Plant, fresh water, soil	Plant rot	1-2(6)	Yes	-	+(2)
*Salmonella enterica**			1(6)	Yes	-	+(1)
Shewanella amazonensis SB2B			1	Yes	-	Yes
*Shewanella baltica**			1(4)	Yes	-	Yes
Shewanella halifaxensis HAW-EB4			2	Yes	-	Yes
Shewanella loihica PV-4			1	Yes	-	Yes
Shewanella oneidensis MR-1			1	Yes	-	Yes
Shewanella pealeana ATCC 700345			2	Yes	-	Yes
*Shewanella sp.**			1(3)	Yes	-	Yes
Shewanella woodyi ATCC 51908			1	Yes	-	Yes

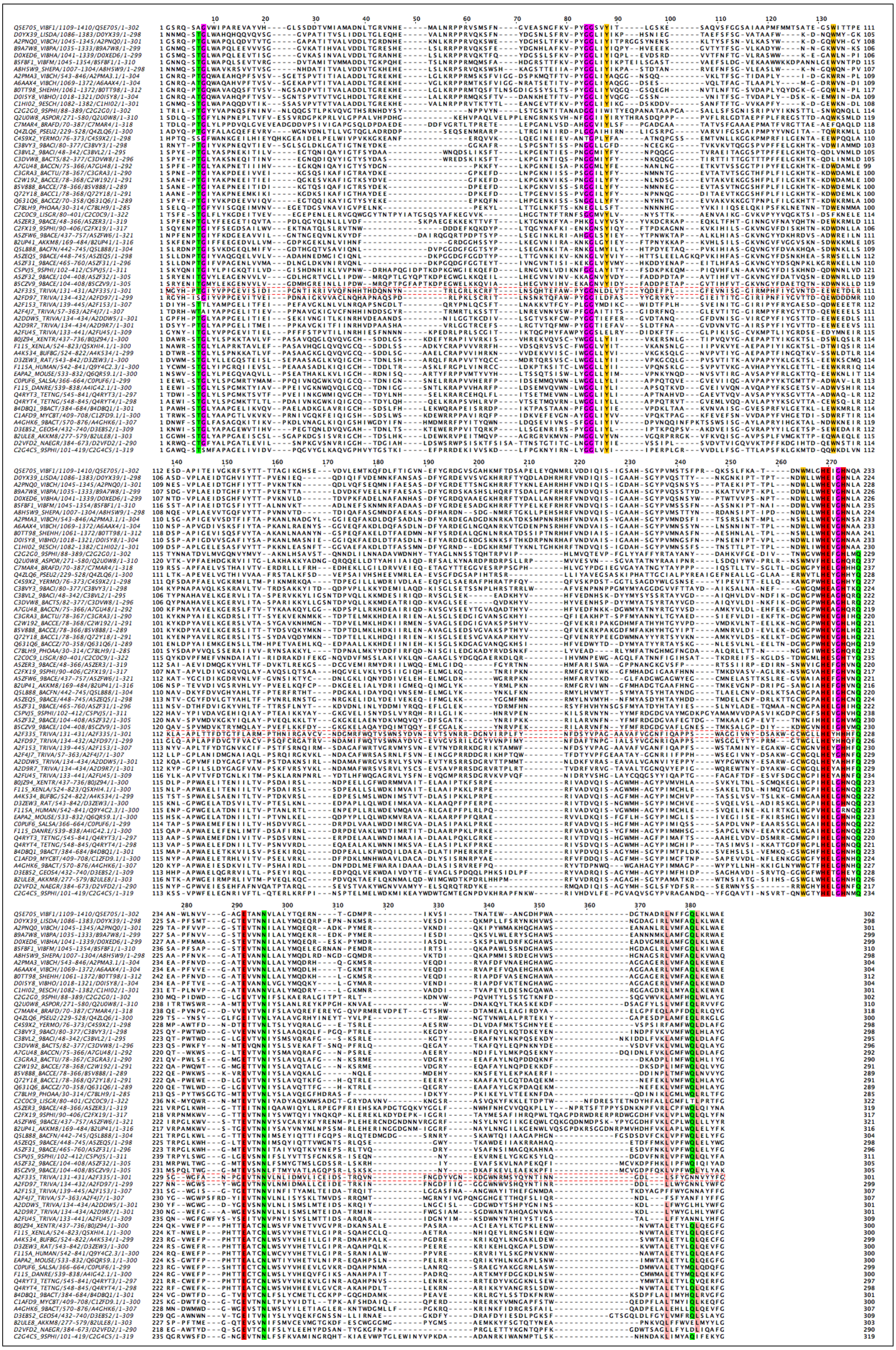
Taxa possessing the M60-like domain	**Habitat or isolation source**	**Disease via or in**	**Number of domain per strain (total per species)**	**HEXXH motif**	**TM**	**SPase**
Shigella sp. D9			1	Yes	-	-
*Vibrio alginolyticus**	Marine	Soft tissue infection	2	Yes	-	Yes
*Vibrio cholerae**	Marine, Fresh water, GIT	GIT	1-2(25)	Yes	-	+(22)
Vibrio fischeri MJ11	Light organs of the squid Euprymna scolopes, Fresh water	-	2	Yes	-	Yes
*Vibrio harveyi**	Marine, GIT	-	2	Yes	-	Yes
*Vibrio mimicus**			1-2(5)	Yes	-	+(4)
Vibrio orientalis CIP 102891	Marine		1	Yes	-	Yes
Vibrio parahaemolyticus 16*		GIT	1-2(12)	Yes	-	+(9)
*Vibrio sp.**			1-2(4)	Yes	-	+(3)
Vibrio splendidus LGP32		Fish and shellfish pathogen	1	Yes	-	Yes
*Vibrio vulnificus**	-	GIT	1-2(3)	Yes	-	+(2)
Vibrionales bacterium SWAT-3	Marine		1	Yes	-	Yes
Yersinia aldovae ATCC 35236	Water	-	1	Yes	-	Yes
Yersinia enterocolitica subsp. enterocolitica 8081	Host	GIT	1	Yes	-	Yes
Yersinia kristensenii ATCC 33638	Environment	?	1	Yes	-	-
Yersinia mollaretii ATCC 43969	Fresh water, Host, Soil	GIT	1	Yes	-	Yes
*Yersinia pestis**		Plague	1-2(21)	Yes	-	-
*Yersinia pseudotuberculosis**	Water	GIT, RT	1(4)	Yes	-	-
Yersinia ruckeri ATCC 29473	Fish with enteric red mouth disease	Fish eteric disease	1	Yes	-	Yes
Planctomycetes						
Planctomyces limnophilus DSM 3776	Surface water	-	1	Yes	-	Yes
Verrucomicrobia						
Akkermansia muciniphila ATCC BAA-835	GIT	-	4	Yes	+(1)	+(2)
Chthoniobacter flavus Ellin428	Soil	-	1	Yes	-	Yes
Verrucomicrobium spinosum DSM 4136	Soil	-	1	Yes	-	Yes
Virus						
Baculoviridae						
Choristoneura fumiferana MNPV			1	Yes	Yes	-
Euproctis pseudoconspersa nucleopolyhedrovirus			1	Yes	Yes	-
Helicoverpa armigera granulovirus			2	Yes	-	-
Lymantria dispar MNPV			2	Yes	-	Yes
Mamestra configurata NPV-A			1	Yes	Yes	-
Xestia c-nigrum granulovirus			2	Yes	-	-
Eukaryota						
Fungi						
Aspergillus flavus NRRL3357	-	RT, plant	2	Yes	-	+(1)
Aspergillus oryzae RIB40	Used in fermented food production	-	2	Yes	-	+(1)
Uncinocarpus reesii 1704	free living	-	1	Yes	-	-
Apicomplexan						
Cryptosporidium muris RN66	-	GIT	1	Yes	-	-
Cryptosporidium parvum Iowa II	GIT	GIT	1	Yes	Yes	-
Amoebozoa						
Entamoeba dispar SAW760	GIT	-	1	Yes	-	Yes
Entamoeba histolytica HM-1:IMSS	GIT	GIT	1	Yes	-	Yes
Choanoflagellida						
Monosiga brevicollis MX1	Marine	-	1	Yes	-	-
Pararasalidea						
Trichomonas vaginalis G3	Host	UGT	25	+(11)	+(6)	+(4)
Mammals						

Taxa possessing the M60-like domain	**Habitat or isolation source**	**Disease via or in**	**Number of domain per strain (total per species)**	**HEXXH motif**	**TM**	**SPase**
Homo sapiens (Human)			5	+(3)	-	-
Pan troglodytes			2	+(1)	-	-
Pongo abelii			1	-	-	-
Macaca mulatta (Rhesus monkey)			1	Yes	-	Yes
Bos taurus			2	+(1)	-	-
Equus caballus			2	+(1)	-	-
Canis familiaris			3	+(2)	-	-
Mus musculus			3	+(2)	-	-
Rattus norvegicus			5	+(3)	-	-
Ornithorhynchus anatinus			3	Yes	-	-
Birds						
Taeniopygia guttata (Zebra finch)			2	+(1)	-	-
Amphibians						
Xenopus laevis (African clawed fog)			1	Yes	-	-
Fish						
Danio rerio			4	+(3)	-	-
Cephalochodata						
Branchiostoma floridae			2	+(1)	-	-

Appendix K

Multiple sequence alignment of the M60-like domains

Proteins are represented by their UniProt accessions and domain regions. The colouring reflects the consensus above 80% conservation. The consensus residues are coloured according to their physico-chemical properties: magenta is Proline or Glycine (conformational special); orange is aromatic; red is positively or negatively charged; green is hydrophilic.



Appendix L

MEROPS proteases possessing M60-like domains

The table contains 38 MEROPS entries that were hit with M60-like domain (PF13402) using HMMER search with cut-off e-value < 1 $\times 10^{-5}$. For each entry, MEROPS identifiers with their descriptions, scores and e-values are shown. The M60-enhancin (PF03272) domains are also listed (if present) with their predicted locations. M60-enhancin was identified using InterProScan.

Merops ID	**Merops description**	**M60-like HMMER e-value**	**M60-like HMMER Score**	**M60-like start**	**M60-like end**	**Merops peptidase unit**	**M60-enhacin InterProScan hits**
MER151941	family M60 unassigned peptidases (*Bacillus cereus*) [M60.UPW]	1.10×10^{-91}	310.5	82	377	223-427	not hit
MER150257	family M60 unassigned peptidases (*Bacillus cereus*) [M60.UPW]	2.50×10^{-86}	292.9	75	370	209-420	not hit
MER111749	family M60 unassigned peptidases (*Akkermansia muciniphila*) [M60.UPW]	5.10×10^{-67}	229.5	100	403	250-458	not hit
MER178106	family M60 unassigned peptidases (*Paenibacillus larvae)* [M60.UPW]	7.80×10^{-21}	77.8	624	856	536-870	288-408, 415-526, 532-870
MER191074	enhancin-like peptidase bacteria (*Uncinocarpus reesii*) [M60.003]	4.40×10^{-20}	75.3	102	332	6-780	4-781
MER137019	enhancin-like peptidase bacteria (*Aspergillus oryzae*) [M60.003]	4.90×10^{-20}	75.1	112	327	4-777	1-778
MER162601	enhancin-like peptidase bacteria (*Aspergillus flavus*) [M60.003]	4.90×10^{-20}	75.1	112	327	4-777	1-778
MER095523	enhancin-like peptidase bacteria (*Clostridium botulinum*) [M60.003]	3.90×10^{-19}	72.2	61	296	31-788	29-792
MER125977	family M60 unassigned peptidases (*Clostridium perfringens*) [M60.UPW]	4.40×10^{-19}	72	140	364	68-763	72-387
MER066138	enhancin-like peptidase bacteria (*Bacillus cereus*) [M60.003]	2.60×10^{-18}	69.5	137	348	23-543	27-544

Merops ID	Merops description	M60-like HMMER e-value	M60-like HMMER Score	M60-like start	M60-like end	Merops peptidase unit	M60-enhacin Inter-ProScan hits
MER028974	enhancin-like peptidase bacteria (*Bacillus cereus*) [M60.003]	1.20×10^{-17}	67.3	136	344	23-742	27-691
MER042489	enhancin-like peptidase bacteria (*Bacillus thuringiensis*) [M60.003]	1.20×10^{-17}	67.3	136	344	23-742	27-691
MER028844	enhancin-like peptidase bacteria (*Bacillus anthracis*) [M60.003]	1.50×10^{-17}	67	136	344	23-742	27-691
MER035718	enhancin-like peptidase bacteria (*Yersinia pestis*) [M60.003]	3.30×10^{-17}	65.8	119	328	5-779	9-778
MER042490	enhancin-like peptidase bacteria (*Yersinia pseudotuberculosis*) [M60.003]	3.40×10^{-17}	65.8	114	323	1-774	4-773
MER150255	enhancin-like peptidase bacteria (*Bacillus cereus*) [M60.003]	4.30×10^{-17}	65.5	136	344	23-742	30-691
MER191081	enhancin-like peptidase bacteria (*Yersinia kristensenii*) [M60.003]	1.30×10^{-16}	63.9	120	288	5-306	9-306
MER191079	family M60 unassigned peptidases (*Paenibacillus larvae*) [M60.UPW]	2.40×10^{-16}	63	7	180	1-217	1-202
MER014453	enhancin (*Lymantria dispar nucleopolyhedrovirus*) [M60.001]	9.00×10^{-15}	57.9	101	319	1-782	5-774
MER014457	family M60 unassigned peptidases (*Heliothis armigera* granulovirus) [M60.UPW]	1.50×10^{-14}	57.2	99	324	1-786	1-790
MER136846	family M60 unassigned peptidases (*Helicoverpa armigera granulovirus*) [M60.UPW]	1.50×10^{-14}	57.2	99	324	1-786	1-790
MER125816	family M60 unassigned peptidases (*Helicoverpa armigera granulovirus*) [M60.UPW]	2.30×10^{-14}	56.5	105	326	4-788	8-790
MER014455	family M60 unassigned peptidases (*Xestia C-nigrum* granulovirus) [M60.UPW]	4.90×10^{-14}	55.4	105	326	4-788	8-792
MER014461	family M60 unassigned peptidases (*Xestia C-nigrum* granulovirus) [M60.UPW]	2.70×10^{-13}	53	99	324	1-786	1-790
MER014459	family M60 unassigned peptidases (*Choristoneura fumiferana* granulovirus) [M60.UPW]	2.80×10^{-13}	53	97	323	1-787	1-791
MER014458	family M60 unassigned peptidases (*Pseudaletia unipuncta* granulovirus) [M60.UPW]	8.30×10^{-13}	51.4	104	323	1-787	1-791
MER014456	family M60 unassigned peptidases (*Trichoplusia ni* granulovirus) [M60.UPW]	9.50×10^{-13}	51.2	99	323	1-787	1-791
MER125723	enhancin-like peptidase bacteria (*Salmonella enterica*) [M60.003]	1.40×10^{-12}	50.6	38	284	12-770	3-772
MER034983	family M60 unassigned peptidases (*Choristoneura fumiferana* nuclear polyhedrosis virus) [M60.UPW]	3.50×10^{-12}	49.3	32	316	1-757	2-754
MER191077	enhancin-2 (*Euproctis pseudoconspersa* nucleopolyhedrovirus) [M60.002]	5.90×10^{-11}	45.3	95	283	7-743	7-742
MER014454	enhancin-2 (*Lymantria dispar* nucleopolyhedrovirus) [M60.002]	3.30×10^{-9}	39.6	96	275	1-788	1-788
MER030203	family M60 unassigned peptidases (*Mamestra configurata* nucleopolyhedrovirus) [M60.UPW]	8.90×10^{-8}	34.9	130	290	1-734	2-716

Merops ID	Merops description	M60-like HMMER e-value	M60-like HMMER Score	M60-like start	M60-like end	Merops peptidase unit	M60-enhacin InterProScan hits
MER118974	family M60 unassigned peptidases (*Mamestra configurata* nucleopolyhedrovirus A) [M60.UPW]	9.00×10^{-8}	34.9	130	290	1-734	2-716
MER118934	family M60 unassigned peptidases (*Agrotis segetum* nucleopolyhedrovirus) [M60.UPW]	1.50×10^{-7}	34.1	60	281	7-749	3-737
MER191076	family M60 unassigned peptidases (*Helicoverpa armigera* multiple nucleopolyhedrovirus) [M60.UPW]	1.80×10^{-7}	33.9	126	279	1-734	1-734
MER191073	family M60 unassigned peptidases (Mamestra brassicae MNPV) [M60.UPW]	2.00×10^{-7}	33.7	126	279	1-734	1-734
MER029638	family M60 unassigned peptidases (*Mamestra configurata* nucleopolyhedrovirus B) [M60.UPW]	2.30×10^{-7}	33.5	130	279	1-734	1-734
MER191080	enhancin-like peptidase bacteria (Listeria grayi) [M60.003]	1.60×10^{-6}	30.8	7	120	4-509	4-458

Appendix M

Functional categories of COGs and KOGs

INFORMATION STORAGE AND PROCESSING

[J] Translation, ribosomal structure and biogenesis

[A] RNA processing and modification

[K] Transcription

[L] Replication, recombination and repair

[B] Chromatin structure and dynamics

CELLULAR PROCESSES AND SIGNALING

[D] Cell cycle control, cell division, chromosome partitioning

[Y] Nuclear structure

[V] Defense mechanisms

[T] Signal transduction mechanisms

[M] Cell wall/membrane/envelope biogenesis

[N] Cell motility

[Z] Cytoskeleton

[W] Extracellular structures

[U] Intracellular trafficking, secretion, and vesicular transport

[O] Posttranslational modification, protein turnover, chaperones

METABOLISM

[C] Energy production and conversion

[G] Carbohydrate transport and metabolism

[E] Amino acid transport and metabolism

[F] Nucleotide transport and metabolism

[H] Coenzyme transport and metabolism

[I] Lipid transport and metabolism

[P] Inorganic ion transport and metabolism

[Q] Secondary metabolites biosynthesis, transport and catabolism

POORLY CHARACTERIZED

[R] General function prediction only

[S] Function unknown

Bibliography

[Abbott *et al.*, 2008] D. W Abbott, J. M Eirín-López, and A. B Boraston. Insight into ligand diversity and novel biological roles for family 32 carbohydrate-binding modules. *Molecular biology and evolution*, 25(1):155–167, Jan 2008.

[Abergel *et al.*, 2007] C Abergel, V Monchois, D Byrne, S Chenivesse, F Lembo, J.-C Lazzaroni, and J.-M Claverie. Structure and evolution of the ivy protein family, unexpected lysozyme inhibitors in gram-negative bacteria. *Proceedings of the National Academy of Sciences*, 104(15):6394–9, Apr 2007.

[Abraham *et al.*, 1998] S. N Abraham, A. B Jonsson, and S Normark. Fimbriae-mediated host-pathogen cross-talk. *Current Opinion in Microbiology*, 1(1):75–81, Feb 1998.

[Acheson and Luccioli, 2004] D. W. K Acheson and S Luccioli. Microbial-gut interactions in health and disease. mucosal immune responses. *Best practice & research Clinical gastroenterology*, 18(2):387–404, Apr 2004.

[Adl *et al.*, 2005] S. M Adl, A. G. B Simpson, M. A Farmer, R. A Andersen, O. R Anderson, J. R Barta, S. S Bowser, G Brugerolle, R. A Fensome, S Fredericq, T. Y James, S Karpov, P Kugrens, J Krug, C. E Lane, L. A Lewis, J Lodge, D. H Lynn, D. G Mann, R. M McCourt, L Mendoza, O Moestrup, S. E Mozley-Standridge, T. A Nerad, C. A Shearer, A. V Smirnov, F. W Spiegel, and M. F. J. R Taylor. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *The Journal of eukaryotic microbiology*, 52(5):399–451, Jan 2005.

[Ahmed *et al.*, 2007] S Ahmed, G. T Macfarlane, A Fite, A. J McBain, P Gilbert, and S Macfarlane. Mucosa-associated bacterial diversity in relation to human terminal ileum and colonic biopsy samples. *Applied and Environment Microbiology*, 73(22):7435–42, Nov 2007.

[Ahmed, 2009] N Ahmed. A flood of microbial genomes-do we need more? *PLoS ONE*, 4(6):e5831, Jan 2009.

[Albers and Driessen, 2002] S.-V Albers and A. M Driessen. Signal peptides of secreted proteins of the archaeon sulfolobus solfataricus: a genomic survey. *Archives of microbiology*, 177(3):209–216, Mar 2002.

[Alberts *et al.*, 2007] B Alberts, A Johnson, J Lewis, M Raff, K Roberts, and P Walter. Molecular biology of the cell (fifth edition). *Garland Science Publishing*, 2007.

[Althouse *et al.*, 2003] C Althouse, S Patterson, P Fedorka-Cray, and R. E Isaacson. Type 1 fimbriae of salmonella enterica serovar typhimurium bind to enterocytes and contribute to colonization of swine in vivo. *Infection and Immunity*, 71(11):6446–52, Nov 2003.

[Anderson and OToole, 2008] G Anderson and G. A OToole. Bacterial biofilms: Innate and induced resistance mechanisms of bacterial biofilms. *Springer-Verlag Berlin Heidelberg*, Current Topics in Microbiology and Immunology 322:85–102, Jan 2008.

[Andersson, 2009] J. O Andersson. Gene transfer and diversification of microbial eukaryotes. *Annual review of microbiology*, 63:177–193, Jan 2009.

[Andrade *et al.*, 2002] M. A Andrade, F. D Ciccarelli, C Perez-Iratxeta, and P Bork. Neat: a domain duplicated in genes near the components of a putative fe3+ siderophore transporter from gram-positive pathogenic bacteria. *Genome Biology*, 3(9):research0047–research0047.5, Aug 2002.

[Andrade *et al.*, 2006] J Andrade, L Berglund, M Uhlén, and J Odeberg. Using grid technology for computationally intensive applied bioinformatics analyses. *In Silico Biology*, 6(6):495–504, 2006.

[Antúnez *et al.*, 2009] K Antúnez, M Anido, G Schlapp, J. D Evans, and P Zunino. Characterization of secreted proteases of paenibacillus larvae, potential virulence factors involved in honeybee larval infection. *Journal of invertebrate pathology*, 102(2):129–132, Oct 2009.

[Arnesen *et al.*, 2008] S Arnesen, P Lotte, A Fagerlund, and P Granum. From soil to gut: Bacillus cereus and its food poisoning toxins. *FEMS Microbiology Reviews*, 32:576–606, Jan 2008.

[Arnold *et al.*, 2009] R Arnold, S Brandmaier, F Kleine, P Tischler, E Heinz, S Behrens, A Niinikoski, H.-W Mewes, M Horn, and T Rattei. Sequence-based prediction of type iii secreted proteins. *PLoS Pathogens*, 5(4):e1000376, Apr 2009.

[Ashburner *et al.*, 2000] M Ashburner, C Ball, J Blake, and D Botstein. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, Jan 2000.

[Attwood, 2002] T. K Attwood. The prints database: a resource for identification of protein families. *Briefings in Bioinformatics*, 3(3):252–263, Sep 2002.

[Audic *et al.*, 2007] S Audic, C Robert, B Campagna, H Parinello, J.-M Claverie, D Raoult, and M Drancourt. Genome analysis of minibacterium massiliensis highlights the convergent evolution of water-living bacteria. *PLoS Genetics*, 3(8):e138, Jan 2007.

[Bäckhed *et al.*, 2005] F Bäckhed, R. E Ley, J. L Sonnenburg, D. A Peterson, and J. I Gordon. Host-bacterial mutualism in the human intestine. *Science*, 307(5717):1915–20, Mar 2005.

[Bae and Schneewind, 2003] T Bae and O Schneewind. The ysirk-g/s motif of staphylococcal protein a and its role in efficiency of signal peptide processing. *Journal of Bacteriology*, 185(9):2910–9, May 2003.

[Bagos *et al.*, 2004a] P. G Bagos, T. D Liakopoulos, I. C Spyropoulos, and S. J Hamodrakas. A hidden markov model method, capable of predicting and discriminating beta-barrel outer membrane proteins. *BMC Bioinformatics*, 5:29, Mar 2004.

[Bagos *et al.*, 2004b] P. G Bagos, T. D Liakopoulos, I. C Spyropoulos, and S. J Hamodrakas. Pred-tmbb: a web server for predicting the topology of beta-barrel outer membrane proteins. *Nucleic Acids Research*, 32(Web Server issue):W400–4, Jul 2004.

[Bagos *et al.*, 2009] P. G Bagos, K. D Tsirigos, S. K Plessas, T. D Liakopoulos, and S. J Hamodrakas. Prediction of signal peptides in archaea. *Protein Engineering Design and Selection*, 22(1):27–35, Jan 2009.

[Baker *et al.*, 2002] M Baker, R Buyya, and D Laforenza. Software: Practice and experience - grids and grid technologies for wide-area distributed computing. *John Wiley & Sons, Inc.*, 32(15):1437–66, Jan 2002.

[Baldock and Burger, 2005] R Baldock and A Burger. Anatomical ontologies: names and places in biology. *Genome Biology*, 6(4):108, Jan 2005.

[Barczak and Hung, 2009] A. K Barczak and D. T Hung. Productive steps toward an antimicrobial targeting virulence. *Current Opinion in Microbiology*, 12(5):490–496, Oct 2009.

[Bardy *et al.*, 2003] S. L Bardy, J Eichler, and K. F Jarrell. Archaeal signal peptides–a comparative survey at the genome level. *Protein Science*, 12(9):1833–43, Sep 2003.

[Barker *et al.*, 1999] W. C Barker, J. S Garavelli, P. B McGarvey, C. R Marzec, B. C Orcutt, G. Y Srinivasarao, L. S Yeh, R. S Ledley, H. W Mewes, F Pfeiffer, A Tsugita, and C Wu. The pir-international protein sequence database. *Nucleic Acids Research*, 27(1):39–43, Jan 1999.

[Bateman, 2010] A Bateman. *Personal communication*, wellcome trust sanger institute. 2010.

[Bearson *et al.*, 1998] B. L Bearson, L Wilson, and J. W Foster. A low ph-inducible, phopq-dependent acid tolerance response protects salmonella typhimurium against inorganic acid stress. *Journal of Bacteriology*, 180(9):2409–17, May 1998.

[Bellgard *et al.*, 2009] M. I Bellgard, P Wanchanthuek, T La, K Ryan, P Moolhuijzen, Z Albertyn, B Shaban, Y Motro, D. S Dunn, D Schibeci, A Hunter, R Barrero, N. D Phillips, and D. J Hampson. Genome sequence of the pathogenic intestinal spirochete brachyspira hyodysenteriae reveals adaptations to its lifestyle in the porcine large intestine. *PLoS ONE*, 4(3):e4641, Jan 2009.

[Bendtsen *et al.*, 2005a] J Bendtsen, T Binnewies, and P Hallin. Genome update: prediction of membrane proteins in prokaryotic genomes. *Microbiology*, 151(7):2119–21, Jan 2005.

[Bendtsen *et al.*, 2005b] J Bendtsen, L Kiemer, A Fausboll, and S Brunak. Non-classical protein secretion in bacteria. *BMC Microbiology*, 5(58), Jan 2005.

[Bendtsen, 2005] J. D Bendtsen. Genome update: prediction of secreted proteins in 225 bacterial proteomes. *Microbiology*, 151(6):1725–7, Jun 2005.

[Benson *et al.*, 2009] D. A Benson, I Karsch-Mizrachi, D. J Lipman, J Ostell, and E. W Sayers. Genbank. *Nucleic Acids Research*, 37(Database issue):D26–31, Jan 2009.

[Benson *et al.*, 2010] A Benson, S Kelly, R Legge, F Ma, S. J Low, J Kim, M Zhang, P. L Oh, D Nehrenberg, K Hua, S. D Kachman, E. N Moriyama, J Walter, D. A Peterson, and D Pomp. Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proceedings of the National Academy of Sciences*, 107(44):18933–38, Jan 2010.

[Bhattacharyya *et al.*, 1988] S. N Bhattacharyya, B Kaufman, A Khorrami, J. I Enriquez, and B Manna. Fibronectin: source of mannose in a highly purified respiratory mucin. *Inflammation*, 12(5):433–446, Oct 1988.

[Bielaszewska and Karch, 2005] M Bielaszewska and H Karch. Consequences of enterohaemorrhagic escherichia coli infection for the vascular endothelium. *Thromb Haemost*, 94(2):312–318, Aug 2005.

[Billion *et al.*, 2006] A Billion, R Ghai, T Chakraborty, and T Hain. Augur–a computational pipeline for whole genome microbial surface protein prediction and classification. *Bioinformatics*, 22(22):2819–20, Sep 2006.

[Bingle *et al.*, 2008] L Bingle, C Bailey, and M Pallen. Type vi secretion: a beginner's guide. *Current Opinion in Microbiology*, 11(1):3–8, Feb 2008.

[Blattner, 1997] F. R Blattner. The complete genome sequence of escherichia coli k-12. *Science*, 277(5331):1453–62, Sep 1997.

[Blum and Schiffrin, 2003] S Blum and E. J Schiffrin. Intestinal microflora and homeostasis of the mucosal immune response: implications for probiotic bacteria? *Current issues in intestinal microbiology*, 4(2):53–60, Sep 2003.

[Bode *et al.*, 1993] W Bode, F Gomis-Rüth, and W Stöckler. Astacins, serralysins, snake venom and matrix metalloproteinases exhibit identical zinc-binding environments (hexxhxxgxxh and met-turn) and topologies and should be grouped into a common family, the 'metzincins'. *FEBS letters*, 331(1):134–140, Jan 1993.

[Boekhorst *et al.*, 2006] J Boekhorst, Q Helmer, M Kleerebezem, and R. J Siezen. Comparative analysis of proteins with a mucus-binding domain found exclusively in lactic acid bacteria. *Microbiology (Reading, Engl)*, 152:273–280, Jan 2006.

[Boraston *et al.*, 2004] A Boraston, D Bolam, and H Gilbert. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochemistry Journal*, 384:769–781, Jan 2004.

[Bouguenec, 2005] C Bouguenec. Adhesins and invasins of pathogenic escherichia coli. *International Journal of Medical Microbiology*, 295(6-7):471–478, Oct 2005.

[Bramley and Kornberg, 1987] H. F Bramley and H. L Kornberg. Sequence homologies between proteins of bacterial phosphoenolpyruvate-dependent sugar phosphotransferase systems: identification of possible phosphate-carrying histidine residues. *Proceedings of the National Academy of Sciences*, 84(14):4777–80, Jul 1987.

[Brun *et al.*, 1997] E Brun, F Moriaud, P Gans, and M Blackledge. Solution structure of the cellulose-binding domain of the endoglucanase z secreted by erwinia chrysanthemi. *Biochemistry*, 36:16074–86, Jan 1997.

[Brynestad and Granum, 2002] S Brynestad and P. E Granum. Clostridium perfringens and foodborne infections. *International journal of food microbiology*, 74(3):195–202, Apr 2002.

[Bunikis *et al.*, 1995] J Bunikis, L Noppa, and S Bergström. Molecular analysis of a 66-kda protein associated with the outer membrane of lyme disease borrelia. *FEMS Microbiology Letters*, 131(2):139–145, Sep 1995.

[Bütikofer *et al.*, 2001] P Bütikofer, T Malherbe, M Boschung, and I Roditi. Gpi-anchored proteins: now you see 'em, now you don't. *FASEB Journal*, 15(2):545–548, Feb 2001.

[Buts *et al.*, 2003] L Buts, J Bouckaert, E. D Genst, R Loris, S Oscarson, M Lahmann, J Messens, E Brosens, L Wyns, and H. D Greve. The fimbrial adhesin f17-g of enterotoxigenic escherichia coli has an immunoglobulin-like lectin domain that binds n-acetylglucosamine. *Molecular Microbiology*, 49(3):705–715, Aug 2003.

[Cabanes *et al.*, 2002] D Cabanes, P Dehoux, O Dussurget, and L Frangeul. Surface proteins and the pathogenic potential of listeria monocytogenes. *Trends in Microbiology*, 10(5):238–245, Jan 2002.

[Carlton *et al.*, 2007] J. M Carlton, R. P Hirt, J. C Silva, A. L Delcher, M Schatz, Q Zhao, J. R Wortman, S. L Bidwell, U. C. M Alsmark, S Besteiro, T Sicheritz-Ponten, C. J Noel, J. B Dacks, P. G Foster, C Simillion, Y. V. D Peer, D Miranda-Saavedra, G. J Barton, G. D Westrop, S Muller, D Dessi, P. L Fiori, Q Ren, I Paulsen, H Zhang, F. D Bastida-Corcuera, A Simoes-Barbosa, M. T Brown, R. D Hayes, M Mukherjee, C. Y Okumura, R Schneider, A. J Smith, S Vanacova, M Villalvazo, B. J Haas, M Pertea, T. V Feldblyum, T. R Utterback, C.-L Shu, K Osoegawa, P. J. D Jong, I Hrdy, L Horvathova, Z Zubacova, P Dolezal, S.-B Malik, J. M Logsdon, K Henze, A Gupta, C. C Wang, R. L Dunne, J. A Upcroft, P Upcroft, O White, S. L Salzberg, P Tang, C.-H Chiu, Y.-S Lee, T. M Embley, G. H Coombs, J. C Mottram, J Tachezy, C. M Fraser-Liggett, and P. J Johnson. Draft genome sequence of the sexually transmitted pathogen trichomonas vaginalis. *Science*, 315(5809):207–212, Jan 2007.

[Casadevall and Pirofski, 2001] A Casadevall and L Pirofski. Host-pathogen interactions: the attributes of virulence. *The Journal of infectious diseases*, 184(3):337–344, Aug 2001.

[Casadio *et al.*, 2008] R Casadio, P Martelli, and A Pierleoni. The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Briefings in Functional Genomics and Proteomics*, 7(1):63–73, Jan 2008.

[Cases *et al.*, 2003] I Cases, V d Lorenzo, and C Ouzounis. Transcription regulation and environmental adaptation in bacteria. *Trends in Microbiology*, 11(248-253), Jan 2003.

[Cegelski *et al.*, 2008] L Cegelski, G. R Marshall, G. R Eldridge, and S. J Hultgren. The biology and future prospects of antivirulence therapies. *Nature Reviews Microbiology*, 6(1):17–27, Jan 2008.

[Chapman *et al.*, 2006] T Chapman, X Wu, and I Barchia. Comparison of virulence gene profiles of escherichia coli strains isolated from healthy and diarrheic swine. *Applied and environmental microbiology*, 72(7):4782–95, Jan 2006.

[Chen *et al.*, 2009] S. L Chen, C. S Hung, J. S Pinkner, J. N Walker, C. K Cusumano, Z Li, J Bouckaert, J. I Gordon, and S. J Hultgren. Positive selection identifies an in vivo role for fimh during urinary tract infection in addition to mannose binding. *Proceedings of the National Academy of Sciences*, 106(52):22439–44, Dec 2009.

[Cho and Salyers, 2001] K. H Cho and A. A Salyers. Biochemical analysis of interactions between outer membrane proteins that contribute to starch utilization by bacteroides thetaiotaomicron. *Journal of Bacteriology*, 183(24):7224–30, Dec 2001.

[Cho *et al.*, 2006] U. S Cho, M. W Bader, M. F Amaya, M. E Daley, R. E Klevit, S. I Miller, and W Xu. Metal bridges between the phoq sensor domain and the membrane regulate transmembrane signaling. *Journal of Molecular Biology*, 356(5):1193–206, Mar 2006.

[Ciccarelli *et al.*, 2006] F. D Ciccarelli, T Doerks, C v Mering, C. J Creevey, B Snel, and P Bork. Toward automatic reconstruction of a highly resolved tree of life. *Science*, 311(5765):1283–7, Mar 2006.

[Clatworthy *et al.*, 2007] A. E Clatworthy, E Pierson, and D. T Hung. Targeting virulence: a new paradigm for antimicrobial therapy. *Nature chemical biology*, 3(9):541–548, Sep 2007.

[Cohen and Hersh, 2005] A Cohen and W Hersh. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71, Jan 2005.

[Cohen and Hunter, 2008] K. B Cohen and L Hunter. Getting started in text mining. *PLoS Computational Biology*, 4(1):e20, Jan 2008.

[Cohen, 2003] B Cohen. Incentives build robustness in bittorrent. *In Proceedings of the 1st Workshop on Economics of Peer-to-Peer Systems*, Jan 2003.

[Consortium *et al.*, 2010] H. M. J. R. S Consortium, K. E Nelson, G. M Weinstock, S. K Highlander, K. C Worley, H. H Creasy, J. R Wortman, D. B Rusch, M Mitreva, E Sodergren, A. T Chinwalla, M Feldgarden, D Gevers, B. J Haas, R Madupu, D. V Ward, B. W Birren, R. A Gibbs, B Methe,

J. F Petrosino, R. L Strausberg, G. G Sutton, O. R White, R. K Wilson, S Durkin, M. G Giglio, S Gujja, C Howarth, C. D Kodira, N Kyrpides, T Mehta, D. M Muzny, M Pearson, K Pepin, A Pati, X Qin, C Yandava, Q Zeng, L Zhang, A. M Berlin, L Chen, T. A Hepburn, J Johnson, J McCorrison, J Miller, P Minx, C Nusbaum, C Russ, S. M Sykes, C. M Tomlinson, S Young, W. C Warren, J Badger, J Crabtree, V. M Markowitz, J Orvis, A Cree, S Ferriera, L. L Fulton, R. S Fulton, M Gillis, L. D Hemphill, V Joshi, C Kovar, M Torralba, K. A Wetterstrand, A Abouelleil, A. M Wollam, C. J Buhay, Y Ding, S Dugan, M. G FitzGerald, M Holder, J Hostetler, S. W Clifton, E Allen-Vercoe, A. M Earl, C. N Farmer, K Liolios, M. G Surette, Q Xu, C Pohl, K Wilczek-Boney, and D Zhu. A catalog of reference genomes from the human microbiome. *Science*, 328(5981):994–999, May 2010.

[Cornell *et al.*, 2007] M Cornell, I Alam, D Soanes, and H Wong. Comparative genome analysis across a kingdom of eukaryotic organisms: Specialization and diversification in the fungi. *Genome Research*, 17:1809–22, Jan 2007.

[Corpet *et al.*, 1999] F Corpet, J Gouzy, and D Kahn. Recent improvements of the prodom database of protein domain families. *Nucleic Acids Research*, 27(1):263–267, Jan 1999.

[Costello *et al.*, 2009] E. K Costello, C. L Lauber, M Hamady, N Fierer, J. I Gordon, and R Knight. Bacterial community variation in human body habitats across space and time. *Science*, 326(5960):1694–7, Dec 2009.

[Cover and Thomas, 1991] T Cover and J Thomas. Elements of information theory. *Wiley Online Library*, Jan 1991.

[Craddock *et al.*, 2008] T Craddock, C. R Harwood, J Hallinan, and A Wipat. e-science: relieving bottlenecks in large-scale genome analyses. *Nature Reviews Microbiology*, 6(12):948–54, Dec 2008.

[Cristóbal *et al.*, 1999] S Cristóbal, J. W d Gier, H Nielsen, and G v Heijne. Competition between sec- and tat-dependent protein translocation in escherichia coli. *The EMBO Journal*, 18(11):2982–90, Jun 1999.

[Cullen, 2004] P Cullen. Outer membrane proteins of pathogenic spirochetes. *FEMS Microbiology Reviews*, 28(3):291–318, Jun 2004.

[Curtis *et al.*, 1999] M. A Curtis, S. A Hanley, and J Aduse-Opoku. The rag locus of porphyromonas gingivalis: a novel pathogenicity island. *Journal of periodontal research*, 34(7):400–5, Oct 1999.

[de Hoon *et al.*, 2004] M. J. L d Hoon, S Imoto, J Nolan, and S Miyano. Open source clustering software. *Bioinformatics*, 20(9):1453–4, Jun 2004.

[de Nobel *et al.*, 2000] H d Nobel, H v. d Ende, and F Klis. Cell wall maintenance in fungi. *Trends in Microbiology*, 8(8):344–345, Jan 2000.

[Decker *et al.*, 2001] K Decker, X Zheng, and C Schmidt. A multi-agent system for automated genomic annotation. *Proceedings of the fifth international conference on Autonomous agents, ACM*, pages 433–440, Jan 2001.

[Deng *et al.*, 2004] W Deng, J Puente, S Gruenheid, Y Li, and B Vallance. Dissecting virulence: Systematic and functional analyses of a pathogenicity island. *Proceedings of the National Academy of Sciences*, 101(10):3597–602, Jan 2004.

[Desvaux *et al.*, 2006] M Desvaux, E Dumas, I Chafsey, and M Hebraud. Protein cell surface display in gram-positive bacteria: from single protein to macromolecular protein structure. *FEMS Microbiology Letters*, pages 1–15, Jan 2006.

[Dethlefsen *et al.*, 2007] L Dethlefsen, M Mcfall-Ngai, and D. A Relman. An ecological and evolutionary perspective on human–microbe mutualism and disease. *Nature*, 449(7164):811–818, Oct 2007.

[Dietrich *et al.*, 2003] G Dietrich, S Kurz, C Hübner, C Aepinus, S Theiss, M Guckenberger, U Panzner, J Weber, and M Frosch. Transcriptome analysis of neisseria meningitidis during infection. *Journal of Bacteriology*, 185(1):155–164, Jan 2003.

[Dinges *et al.*, 2000] M. M Dinges, P. M Orwin, and P. M Schlievert. Exotoxins of staphylococcus aureus. *Clinical microbiology reviews*, 13(1):16–34, table of contents, Jan 2000.

[Dodson *et al.*, 2001] K. W Dodson, J. S Pinkner, T Rose, G Magnusson, S. J Hultgren, and G Waksman. Structural basis of the interaction of the pyelonephritic e. coli adhesin to its human kidney receptor. *Cell*, 105(6):733–743, Jun 2001.

[Dreisbach *et al.*, 2010] A Dreisbach, K Hempel, G Buist, M Hecker, D Becher, and J. M v Dijl. Profiling the surfacome of staphylococcus aureus. *Proteomics*, 10(17):3082–96, Sep 2010.

[Duncan *et al.*, 2007] S. H Duncan, A Belenguer, G Holtrop, A. M Johnstone, H. J Flint, and G. E Lobley. Reduced dietary intake of carbohydrates by obese subjects results in decreased concentrations of butyrate and butyrate-producing bacteria in feces. *Applied and Environment Microbiology*, 73(4):1073–8, Feb 2007.

[Dutta and Pan, 2002] C Dutta and A Pan. Horizontal gene transfer and bacterial diversity. *Journal of biosciences*, 27(1 Suppl 1):27–33, Feb 2002.

[Dyrlovbendtsen, 2004] J Dyrlovbendtsen. Improved prediction of signal peptides: Signalp 3.0. *Journal of Molecular Biology*, 340(4):783–795, Jul 2004.

[Eddy, 1998] S Eddy. Profile hidden markov models. *Bioinformatics*, 14(9):755–763, Jan 1998.

[Edman *et al.*, 1990] U Edman, M Meraz, S Rausser, and N Agabian. Characterization of an immuno-dominant variable surface antigen from pathogenic and nonpathogenic entamoeba histolytica. *Journal of Experimental Medicine*, 172:879–888, Jan 1990.

[Eichler and Adams, 2005] J Eichler and M. W. W Adams. Posttranslational protein modification in archaea. *Microbiology and molecular biology reviews*, 69(3):393–425, Sep 2005.

[Eisenhaber *et al.*, 1999] B Eisenhaber, P Bork, and F Eisenhaber. Prediction of potential gpi-modification sites in proprotein sequences. *Journal of Molecular Biology*, 292(3):741–58, Sep 1999.

[Eisenhaber *et al.*, 2000] B Eisenhaber, P Bork, Y Yuan, G Löffler, and F Eisenhaber. Automated annotation of gpi anchor sites: case study c. elegans. *Trends in Biochemical Sciences*, 25(7):340–341, Jul 2000.

[Ellen *et al.*, 2010] A. F Ellen, B Zolghadr, A. M. J Driessen, and S.-V Albers. Shaping the archaeal cell envelope. *Archaea*, 2010:608243, Jan 2010.

[Ellrott *et al.*, 2010] K Ellrott, L Jaroszewski, W Li, J. C Wooley, and A Godzik. Expansion of the protein repertoire in newly explored environments: human gut microbiome specific protein families. *PLoS Computational Biology*, 6(6):e1000798, Jan 2010.

[Faith, 2007] J Faith. Targeted projection pursuit for interactive exploration of high-dimensional data sets. *Proceedings of the 11th International Conference Information Visualization*, IEEE Computer Society:286–292, Jan 2007.

[Farmer *et al.*, 1985] J. J Farmer, G. R Fanning, B. R Davis, C. M O'Hara, C Riddle, F. W Hickman-Brenner, M. A Asbury, V. A Lowery, and D. J Brenner. Escherichia fergusonii and enterobacter taylorae, two new species of enterobacteriaceae isolated from clinical specimens. *Journal of clinical microbiology*, 21(1):77–81, Jan 1985.

[Fedhila *et al.*, 2006] S Fedhila, N Daou, D Lereclus, and C Nielsen-LeRoux. Identification of bacillus cereus internalin and other candidate virulence genes specifically induced during oral infection in insects. *Molecular Microbiology*, 62(2):339–55, Oct 2006.

[Field *et al.*, 2008] D Field, G Garrity, T Gray, N Morrison, J Selengut, P Sterk, T Tatusova, N Thomson, M. J Allen, S. V Angiuoli, M Ashburner, N Axelrod, S Baldauf, S Ballard, J Boore, G Cochrane, J Cole, P Dawyndt, P. D Vos, C dePamphilis, R Edwards, N Faruque, R Feldman, J Gilbert, P Gilna, F. O Glöckner, P Goldstein, R Guralnick, D Haft, D Hancock, H Hermjakob, C Hertz-Fowler, P Hugenholtz, I Joint, L Kagan, M Kane, J Kennedy, G Kowalchuk, R Kottmann, E Kolker, S Kravitz, N Kyrpides, J Leebens-Mack, S. E Lewis, K Li, A. L Lister, P Lord, N Maltsev, V Markowitz, J Martiny, B Methe, I Mizrachi, R Moxon, K Nelson, J Parkhill, L Proctor, O White, S.-A Sansone, A Spiers, R Stevens, P Swift, C Taylor, Y Tateno, A Tett, S Turner, D Ussery, B Vaughan, N Ward, T Whetzel, I. S Gil, G Wilson, and A Wipat. The minimum information about a genome sequence (migs) specification. *Nature Biotechnology*, 26(5):541–547, Jan 2008.

[Finn *et al.*, 2010] R. D Finn, J Mistry, J Tate, P Coggill, A Heger, J. E Pollington, O. L Gavin, P Gunasekaran, G Ceric, K Forslund, L Holm, E. L. L Sonnhammer, S. R Eddy, and A Bateman. The pfam protein families database. *Nucleic Acids Research*, 38(Database issue):D211–22, Jan 2010.

[Fitzgerald *et al.*, 2001] J Fitzgerald, D Sturdevant, S. M Mackie, S. R Gill, and J. M Musser. Evolutionary genomics of staphylococcus aureus: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. *Proceedings of the National Academy of Sciences*, 98(15):8821–6, Jan 2001.

[Flanagan, 2009] K Flanagan. A grid and cloud-based framework for high throughput bioinformatics. *PhD thesis - Newcastle University (School of Computing Science)*, 2009.

[Flint *et al.*, 2008] H. J Flint, E. A Bayer, M. T Rincon, R Lamed, and B. A White. Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nature Reviews Microbiology*, 6(2):121–131, Feb 2008.

[Foster and Kesselman, 1997] I Foster and C Kesselman. Globus: A metacomputing infrastructure toolkit. *International journal of supercomputer applications*, pages 1–15, Nov 1997.

[Foster *et al.*, 2001] I Foster, C Kesselman, and S Tuecke. The anatomy of the grid: Enabling scalable virtual organizations. *Proceedings of the 7th International Euro-Par Conference Manchester on Parallel Processing*, pages 1–4, Jan 2001.

[Foster, 2006] I Foster. Globus toolkit version 4: Software for service-oriented systems. *Journal of Computer Science and Technology*, 21(4):513–520, Jun 2006.

[Fouet, 2009] A Fouet. The surface of bacillus anthracis. *Molecular aspects of medicine*, 30(6):374–385, Dec 2009.

[Fraser-Liggett, 2005] C. M Fraser-Liggett. Insights on biology and evolution from microbial genome sequencing. *Genome Research*, 15(12):1603–10, Dec 2005.

[Freeman and Wimley, 2010] T Freeman and W Wimley. A highly accurate statistical approach for the prediction of transmembrane beta-barrels. *Bioinformatics*, 26(16):1965–74, Jun 2010.

[Frey *et al.*, 2002] J Frey, T Tannenbaum, M Livny, I Foster, and S Tuecke. Condor-g: A computation management agent for multi-institutional grids. *Cluster Computing*, 5:237–246, Jan 2002.

[Frey, 1996] A Frey. Role of the glycocalyx in regulating access of microparticles to apical plasma membranes of intestinal epithelial cells: implications for microbial attachment and oral vaccine targeting. *Journal of Experimental Medicine*, 184(3):1045–59, Jan 1996.

[Friedman *et al.*, 2001] C Friedman, P Kra, H Yu, M Krauthammer, and A Rzhetsky. Genies: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(1):S74–S82, Jan 2001.

[Garcia-Vallve *et al.*, 2003] S Garcia-Vallve, E Guzman, M. A Montero, and A Romeu. Hgt-db: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Research*, 31(1):187–189, Jan 2003.

[Gardner *et al.*, 2006] M Gardner, W Feng, J Archuleta, H Lin, and X Mal. Parallel genomic sequence-searching on an ad-hoc grid: Experiences, lessons learned, and implications. *Proceedings of the 2006 CM/IEEE conference on Supercomputing*, page 22, Jan 2006.

[Gardy and Brinkman, 2006] J. L Gardy and F. S. L Brinkman. Methods for predicting bacterial protein subcellular localization. *Nature Reviews Microbiology*, 4(10):741–751, Oct 2006.

[Gardy, 2004] J. L Gardy. Psortb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*, 21(5):617–623, Sep 2004.

[Gill *et al.*, 2006] S Gill, M Pop, R Deboy, P Eckburg, P Turnbaugh, B Samuel, J Gordon, D Relman, C Fraser-Liggett, and K Nelson. Metagenomic analysis of the human distal gut microbiome. *Science*, 312(5778):1355–9, Jun 2006.

[Goh *et al.*, 2006] C.-S Goh, T. A Gianoulis, Y Liu, J Li, A Paccanaro, Y. A Lussier, and M Gerstein. Integration of curated databases to identify genotype-phenotype associations. *BMC Genomics*, 7:257, Jan 2006.

[Golyshina and Timmis, 2005] O Golyshina and K Timmis. Ferroplasma and relatives, recently discovered cell wall-lacking archaea making a living in extremely acid, heavy metal-rich environments. *Environmental Microbiology*, 7(9):1277–88, Jan 2005.

[Gomez *et al.*, 2000] M Gomez, S Johnson, and M Gennaro. Identification of secreted proteins of mycobacterium tuberculosis by a bioinformatic approach. *Infection and Immunity*, 68:2323–7, Jan 2000.

[Gomi *et al.*, 2005a] M Gomi, R Sawada, M.Sonoyama, and S.Mitaku. Comparative proteomics of the prokaryota using secretory proteins. *Chem-Bio Informatics Journal*, 5(3):56–64, Jan 2005.

[Gomi *et al.*, 2005b] M Gomi, M Sonoyama, and S Mitaku. High performance system for signal peptide prediction: Sosuisignal. *Chem-Bio Informatics Journal*, 4(4):142–147, Jan 2005.

[Gomis-Rüth, 2003] F Gomis-Rüth. Structural aspects of the metzincin clan of metalloendopeptidases. *Molecular biotechnology*, 24(2):157–202, Jan 2003.

[Goodman *et al.*, 2009] A. L Goodman, N. P McNulty, Y Zhao, D Leip, R. D Mitra, C. A Lozupone, R Knight, and J. I Gordon. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe*, 6(3):279–289, Sep 2009.

[Gordon *et al.*, 2005] J Gordon, R Ley, R Wilson, E Mardis, and J Xu. Extending our view of self: the human gut microbiome initiative (hgmi). *National Human Genome Research Institute*, Jan 2005.

[Gosset, 2005] G Gosset. Improvement of escherichia coli production strains by modification of the phosphoenolpyruvate:sugar phosphotransferase system. *Microbial Cell Factories*, 4(1):14, May 2005.

[Grigg *et al.*, 2007] J. C Grigg, C. L Vermeiren, D. E Heinrichs, and M. E. P Murphy. Haem recognition by a staphylococcus aureus neat domain. *Molecular Microbiology*, 63(1):139–49, Jan 2007.

[Groth *et al.*, 2008] P Groth, B Weiss, H.-D Pohlenz, and U Leser. Mining phenotypes for gene function prediction. *BMC Bioinformatics*, 9:136, Jan 2008.

[Guénola *et al.*, 2006] R Guénola, M Neil, D Bas, J Jean-Pierre, D Macheboeuf, M Mitsumori, F. M McIntosh, T Michalowski, T Nagamine, N Nelson, C. J Newbold, E Nsabimana, A Takenaka, N. A Thomas, K Ushida, J. H Hackstein, and M. A Huynen. Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. *BMC Genomics*, 7:22, 2006.

[Gupta *et al.*, 2003] A Gupta, H Vlamakis, N Shoemaker, and A. A Salyers. A new bacteroides conjugative transposon that carries an ermb gene. *Applied and Environment Microbiology*, 69(11):6455–63, Jan 2003.

[Haft *et al.*, 2001] D. H Haft, B. J Loftus, D. L Richardson, F Yang, J. A Eisen, I. T Paulsen, and O White. Tigrfams: a protein family resource for the functional identification of proteins. *Nucleic Acids Research*, 29(1):41–43, Jan 2001.

[Han *et al.*, 2006] C. S Han, G Xie, J. F Challacombe, M. R Altherr, S. S Bhotika, N Brown, D Bruce, C. S Campbell, M. L Campbell, J Chen, O Chertkov, C Cleland, M Dimitrijevic, N. A Doggett, J. J Fawcett, T Glavina, L. A Goodwin, L. D Green, K. K Hill, P Hitchcock, P. J Jackson, P Keim, A. R Kewalramani, J Longmire, S Lucas, S Malfatti, K McMurry, L. J Meincke, M Misra, B. L Moseman, M Mundt, A. C Munk, R. T Okinaka, B Parson-Quintana, L. P Reilly, P Richardson, D. L Robinson, E Rubin, E Saunders, R Tapia, J. G Tesmer, N Thayer, L. S Thompson, H Tice, L. O Ticknor, P. L Wills, T. S Brettin, and P Gilna. Pathogenomic sequence analysis of bacillus cereus and bacillus thuringiensis isolates closely related to bacillus anthracis. *Journal of Bacteriology*, 188(9):3382–90, May 2006.

[Hattori and Taylor, 2009] M Hattori and T. D Taylor. The human intestinal microbiome: a new frontier of human biology. *DNA Research*, 16(1):1–12, Feb 2009.

[Heermann and Fuchs, 2008] R Heermann and T. M Fuchs. Comparative analysis of the photorhabdus luminescens and the yersinia enterocolitica genomes: uncovering candidate genes involved in insect pathogenicity. *BMC Genomics*, 9:40, Jan 2008.

[Hehemann *et al.*, 2010] J Hehemann, G Correc, T Barbeyron, W Helbert, M Czjzek, and G Michel. Transfer of carbohydrate-active enzymes from marine bacteria to japanese gut microbiota. *Nature*, 464(7290):908–912, Jan 2010.

[Herrington *et al.*, 1988] D. A Herrington, R. H Hall, G Losonsky, J. J Mekalanos, R. K Taylor, and M. M Levine. Toxin, toxin-coregulated pili, and the toxr regulon are essential for vibrio cholerae pathogenesis in humans. *The Journal of experimental medicine*, 168(4):1487–92, Oct 1988.

[Hirokawa *et al.*, 1998] T Hirokawa, S Boon-Chieng, and S Mitaku. Sosui: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, 14(4):378–379, Jan 1998.

[Hirschman *et al.*, 2008] L Hirschman, C Clark, K. B Cohen, S Mardis, J Luciano, R Kottmann, J Cole, V Markowitz, N Kyrpides, N Morrison, L. M Schriml, D Field, and N Project. Habitat-lite: a gsc case study based on free text terms for environmental metadata. *OMICS*, 12(2):129–136, Jun 2008.

[Hirt *et al.*, 2002] R Hirt, N Harriman, A Kajava, and T Embley. A novel potential surface protein in trichomonas vaginalis contains a leucine-rich repeat shared by micro-organisms from all three domains of life. *Molecular and Biochemical Parasitology*, 125:195–199, Jan 2002.

[Hirt *et al.*, 2007] R Hirt, C Noel, T Sicheritzponten, J Tachezy, and P Fiori. Trichomonas vaginalis surface proteins: a view from the genome. *Trends in Parasitology*, 23(11):540–547, Nov 2007.

[Høiby *et al.*, 2010] N Høiby, T Bjarnsholt, M Givskov, S Molin, and O Ciofu. Antibiotic resistance of bacterial biofilms. *International Journal of AntimicrobialAgents*, 35:322–332, Jan 2010.

[Holden *et al.*, 2004] M Holden, L Crossman, A Cerdeño-Tárraga, and J Parkhill. Pathogenomics of non-pathogens. *Nature Reviews. Microbiology*, 2(2):91–91, Feb 2004.

[Hooper and Gordon, 2001] L. V Hooper and J. I Gordon. Commensal host-bacterial relationships in the gut. *Science*, 292(5519):1115–8, May 2001.

[Hooper, 1994] N. M Hooper. Families of zinc metalloproteases. *FEBS letters*, 354(1):1–6, Oct 1994.

[Houot *et al.*, 2010] L Houot, S Chang, C Absalon, and P. I Watnick. Vibrio cholerae phosphoenolpyruvate phosphotransferase system control of carbohydrate transport, biofilm formation, and colonization of the germfree mouse intestine. *Infection and Immunity*, 78(4):1482–94, Apr 2010.

[Howell *et al.*, 1994] S Howell, C Lanctôt, G Boileau, and P Crine. A cleavable n-terminal signal peptide is not a prerequisite for the biosynthesis of glycosylphosphatidylinositol-anchored proteins. *The Journal of biological chemistry*, 269(25):16993–6, Jun 1994.

[Hsiao, 2003] W Hsiao. Islandpath: aiding detection of genomic islands in prokaryotes. *Bioinformatics*, 19(3):418–420, Feb 2003.

[Hughes *et al.*, 1994] K. J Hughes, K. D Everiss, M. E Kovach, and K. M Peterson. Sequence analysis of the vibrio cholerae acfd gene reveals the presence of an overlapping reading frame, orfz, which encodes a protein that shares sequence similarity to the flia and flic products of salmonella. *Gene*, 146(1):79–82, Aug 1994.

[Hull *et al.*, 2006] D Hull, K Wolstencroft, R Stevens, C Goble, M. R Pocock, P Li, and T Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, 34(Web Server issue):W729–32, Jul 2006.

[Hulo *et al.*, 2006] N Hulo, A Bairoch, V Bulliard, L Cerutti, E d Castro, P. S Langendijk-Genevaux, M Pagni, and C. J. A Sigrist. The prosite database. *Nucleic Acids Research*, 34:227–230, 2006.

[Hung *et al.*, 2002] C.-S Hung, J Bouckaert, D Hung, J Pinkner, C Widberg, A DeFusco, C. G Auguste, R Strouse, S Langermann, G Waksman, and S. J Hultgren. Structural basis of tropism of escherichia coli to the bladder during urinary tract infection. *Molecular Microbiology*, 44(4):903–915, May 2002.

[Hunter *et al.*, 2009] S Hunter, R Apweiler, T Attwood, and A Bairoch. Interpro: the integrative protein signature database. *Nucleic Acids Research*, 37(Database issue):D211–215, Jan 2009.

[Hwang *et al.*, 2002] P. M Hwang, W.-Y Choy, E. I Lo, L Chen, J. D Forman-Kay, C. R. H Raetz, G. G Privé, R. E Bishop, and L. E Kay. Solution structure and dynamics of the outer membrane enzyme pagp by nmr. *Proceedings of the National Academy of Sciences*, 99(21):13560–5, Oct 2002.

[Ibrahim *et al.*, 2008] S Ibrahim, H Jin, L Qi, and C Zeng. Grid maintenance: Challenges and existing models. *Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference*, pages 1–6, Jan 2008.

[Ikegami *et al.*, 2000] T Ikegami, T Okada, M Hashimoto, and S Seino. Solution structure of the chitin-binding domain of bacillus circulans wl-12 chitinase a1. *Journal of Biological Chemistry*, 275(18):13654–61, Jan 2000.

[Isberg and Falkow, 1985] R. R Isberg and S Falkow. A single genetic locus encoded by yersinia pseudotuberculosis permits invasion of cultured animal cells by escherichia coli k-12. *Nature*, 317(6034):262–264, Jan 1985.

[Isberg *et al.*, 1987] R. R Isberg, D. L Voorhis, and S Falkow. Identification of invasin: a protein that allows enteric bacteria to penetrate cultured mammalian cells. *Cell*, 50(5):769–778, Aug 1987.

[Iwase *et al.*, 2010] T Iwase, Y Uehara, H Shinji, A Tajima, H Seo, K Takada, T Agata, and Y Mizunoe. Staphylococcus epidermidis esp inhibits staphylococcus aureus biofilm formation and nasal colonization. *Nature*, 465(7296):346–349, May 2010.

[James *et al.*, 2009] K James, A Wipat, and J Hallinan. Integration of full-coverage probabilistic functional networks with relevance to specific biological processes. *Data Integration in the Life Sciences*, 5647:31–46, Jan 2009.

[Jensen *et al.*, 2006] L. J Jensen, J Saric, and P Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics*, 7(2):119–129, Feb 2006.

[Jia *et al.*, 2008] W Jia, H Li, L Zhao, and J. K Nicholson. Gut microbiota: a potential new territory for drug targeting. *Nature Reviews. Drug Discovery*, 7(2):123–129, Feb 2008.

[Jim, 2003] K Jim. A cross-genomic approach for systematic mapping of phenotypic traits to genes. *Genome Research*, 14(1):109–115, Dec 2003.

[Johnston and Mabey, 2008] V Johnston and D Mabey. Global epidemiology and control of trichomonas vaginalis. *Current Opinion in Infectious Diseases*, 21:56–64, Jan 2008.

[Jones *et al.*, 2008] B. V Jones, M Begley, C Hill, C. G. M Gahan, and J. R Marchesi. Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. *Proceedings of the National Academy of Sciences*, 105(36):13580–5, Sep 2008.

[Jongeneel *et al.*, 1989] C Jongeneel, J Bouvier, and A Bairoch. A unique signature identifies a family of zinc-dependent metallopeptidases. *FEBS letters*, 242(2):211–214, Jan 1989.

[Juncker *et al.*, 2003] A Juncker, H Willenbrock, G v Heijne, and S Brunak. Prediction of lipoprotein signal peptides in gram-negative bacteria. *Protein Science: A Publication of the Protein Society*, 12(8):1652–62, Jan 2003.

[Kall, 2004] L Kall. A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*, 338(5):1027–36, May 2004.

[Kaper, 2005] J Kaper. Pathogenic escherichia coli. *International Journal of Medical Microbiology*, 295:355–356, Jan 2005.

[Karo *et al.*, 2001] M Karo, C Dwan, J Freeman, J Freeman, J Weissman, M Livny, and E Retzel. Applying grid technologies to bioinformatics. *Proceedings of the 10th IEEE International Symposium on High Performance Distributed Computing*, pages 441–442, Jan 2001.

[Keeling and Palmer, 2008] P. J Keeling and J. D Palmer. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8):605–618, Aug 2008.

[Kobe and Kajava, 2001] B Kobe and A Kajava. The leucine-rich repeat as a protein recognition motif. *Current Opinion in Structural Biology*, 11:725–732, Jan 2001.

[Kokoeva *et al.*, 2002] M. V Kokoeva, K.-F Storch, C Klein, and D Oesterhelt. A novel mode of sensory transduction in archaea: binding protein-mediated chemotaxis towards osmoprotectants and amino acids. *The EMBO Journal*, 21(10):2312–22, May 2002.

[Kolstø *et al.*, 2009] A.-B Kolstø, N. J Tourasse, and O. A Økstad. What sets bacillus anthracis apart from other bacillus species? *Annual review of microbiology*, 63:451–476, Jan 2009.

[Korbel *et al.*, 2005] J. O Korbel, T Doerks, L. J Jensen, C Perez-Iratxeta, S Kaczanowski, S. D Hooper, M. A Andrade, and P Bork. Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biology*, 3(5):e134, Jan 2005.

[Krogh *et al.*, 2001] A Krogh, B Larsson, G v Heijne, and E Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology*, 305:567–580, Jan 2001.

[Kuczynski *et al.*, 2010] J Kuczynski, E. K Costello, D. R Nemergut, J Zaneveld, C. L Lauber, D Knights, O Koren, N Fierer, S. T Kelley, R. E Ley, J. I Gordon, and R Knight. Direct sequencing of the human microbiome readily reveals community differences. *Genome Biology*, 11(5):210, May 2010.

[Kurokawa *et al.*, 2007] K Kurokawa, T Itoh, T Kuwahara, K Oshima, H Toh, A Toyoda, H Takami, H Morita, V. K Sharma, T. P Srivastava, T. D Taylor, H Noguchi, H Mori, Y Ogura, D. S Ehrlich, K Itoh, T Takagi, Y Sakaki, T Hayashi, and M Hattori. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Research*, 14(4):169–181, Aug 2007.

[Lao *et al.*, 2002] D. M Lao, M Arai, M Ikeda, and T Shimizu. The presence of signal peptide significantly affects transmembrane topology prediction. *Bioinformatics*, 18(12):1562–6, Dec 2002.

[Lebeis and Kalman, 2009] S. L Lebeis and D Kalman. Aligning antimicrobial drug discovery with complex and redundant host-pathogen interactions. *Cell Host Microbe*, 5(2):114–122, Feb 2009.

[Lee and Schneewind, 2001] V. T Lee and O Schneewind. Protein secretion and the pathogenesis of bacterial infections. *Genes and Development*, 15(14):1725–52, Jul 2001.

[Lee *et al.*, 2003] S. Y Lee, J. H Choi, and Z Xu. Microbial cell-surface display. *Trends in Biotechnology*, 21(1):45–52, Jan 2003.

[Lee *et al.*, 2008] J Lee, V Karamychev, S Kozyavkin, and D Mills. Comparative genomic analysis of the gut bacterium bifidobacterium longum reveals loci susceptible to deletion during pure culture growth. *BMC Genomics*, 9:247, Jan 2008.

[Lepore *et al.*, 1996] L. S Lepore, P. R Roelvink, and R. R Granados. Enhancin, the granulosis virus protein that facilitates nucleopolyhedrovirus (npv) infections, is a metalloprotease. *Journal of invertebrate pathology*, 68(2):131–140, Sep 1996.

[Lerouxel *et al.*, 2006] O Lerouxel, D Cavalier, A Liepman, and K Keegstra. Biosynthesis of plant cell wall polysaccharides—a complex process. *Current Opinion in Plant Biology*, 9:621–630, Jan 2006.

[Levow, 2010] G.-A Levow. *Personal communication*, university of manchester, uk. 2010.

[Ley *et al.*, 2006] R Ley, D Peterson, and J Gordon. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell*, 124:837–848, Jan 2006.

[Ley *et al.*, 2008] R. E Ley, C. A Lozupone, M Hamady, R Knight, and J. I Gordon. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nature Reviews Microbiology*, 6(10):776–788, Oct 2008.

[Li *et al.*, 2003] L Li, C. J Stoeckert, and D. S Roos. Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9):2178–89, Sep 2003.

[Lima *et al.*, 2009] T Lima, A Auchincloss, E Coudert, and G Keller. Hamap: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in uniprotkb/swiss-prot. *Nucleic Acids Research*, 37(Database):D471–8, Jan 2009.

[Lin *et al.*, 2002] J Lin, S Huang, and Q Zhang. Outer membrane proteins: key players for bacterial adaptation in host niches. *Microbes and Infection*, 4(3):325–331, Mar 2002.

[Little and Rubin, 2000] R. J Little and D. B Rubin. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu Rev Public Health*, 21:121–145, Jan 2000.

[Liu *et al.*, 2006] Y Liu, J Li, L Sam, C. S Goh, M Gerstein, and Y. A Lussier. An integrative genomic approach to uncover molecular mechanisms of prokaryotic traits. *PLoS Computational Biology*, 2(11):1419–35, Nov 2006.

[Loftus *et al.*, 2005] B Loftus, I Anderson, R Davies, U. C Alsmark, J Samuelson, P Amedeo, P Roncaglia, M Berriman, R. P Hirt, B. J Mann, T Nozaki, B Suh, M Pop, M Duchene, J Ackers, E Tannich, M Leippe, M Hofer, I Bruchhaus, U Willhoeft, A Bhattacharya, T Chillingworth, C Churcher, Z Hance, B Harris, D Harris, K Jagels, S Moule, K Mungall, D Ormond, R Squares, S Whitehead, M. A Quail, E Rabbinowitsch, H Norbertczak, C Price, Z Wang, N Guillén, C Gilchrist, S. E Stroup, S Bhattacharya, A Lohia, P. G Foster, T Sicheritz-Ponten, C Weber, U Singh, C Mukherjee, N. M El-Sayed, W. A Petri, C. G Clark, T. M Embley, B Barrell, C. M Fraser, and N Hall. The genome of the protist parasite entamoeba histolytica. *Nature*, 433(7028):865–868, Feb 2005.

[Lord *et al.*, 2003] P. W Lord, R. D Stevens, A Brass, and C. A Goble. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–83, Jul 2003.

[Lozupone *et al.*, 2006] C Lozupone, M Hamady, and R Knight. Unifrac–an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics*, 7:371, Jan 2006.

[Lozupone *et al.*, 2008] C Lozupone, M Hamady, B Cantarel, and P Coutinho. The convergence of carbohydrate active gene repertoires in human gut microbes. *Proceedings of the National Academy of Sciences*, 105(39):15076–81, 2008.

[Maere *et al.*, 2005] S Maere, K Heymans, and M Kuiper. Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–9, Aug 2005.

[Marino *et al.*, 2002] M Marino, M Banerjee, R Jonquières, P Cossart, and P Ghosh. Gw domains of the listeria monocytogenes invasion protein inlb are sh3-like and mediate binding to host ligands. *The EMBO Journal*, 21(21):5623–34, Nov 2002.

[Marion and Guillén, 2006] S Marion and N Guillén. Genomic and proteomic approaches highlight phagocytosis of living and apoptotic human cells by the parasite entamoeba histolytica. *International Journal for Parasitology*, 36(2):131–139, Feb 2006.

[Martens *et al.*, 2008] E. C Martens, H. C Chiang, and J. I Gordon. Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host Microbe*, 4(5):447–457, Nov 2008.

[Martens *et al.*, 2009] E. C Martens, N. M Koropatkin, T. J Smith, and J. I Gordon. Complex glycan catabolism by the human gut microbiota: the bacteroidetes sus-like paradigm. *The Journal of biological chemistry*, 284(37):24673–7, Sep 2009.

[Martin *et al.*, 1998] K Martin, G Morlin, A Smith, A Nordyke, A Eisenstark, and M Golomb. The tryptophanase gene cluster of haemophilus influenzae type b: evidence for horizontal gene transfer. *Journal of Bacteriology*, 180(1):107–118, Jan 1998.

[Martin *et al.*, 2007] F.-P. J Martin, M.-E Dumas, Y Wang, C Legido-Quigley, I. K. S Yap, H Tang, S Zirah, G. M Murphy, O Cloarec, J. C Lindon, N Sprenger, L. B Fay, S Kochhar, P. V Bladeren, E Holmes, and J. K Nicholson. A top-down systems biology view of microbiome-mammalian metabolic interactions in a mouse model. *Molecular Systems Biology*, 3:16, May 2007.

[Matsunaga *et al.*, 2009] A Matsunaga, M Tsugawa, and J Fortes. Cloudblast: Combining mapreduce and virtualization on distributed resources for bioinformatics applications. *Fourth IEEE International Conference on eScience*, pages 222–229, Jan 2009.

[Mattar *et al.*, 1994] S Mattar, B Scharf, S. B Kent, K Rodewald, D Oesterhelt, and M Engelhard. The primary structure of halocyanin, an archaeal blue copper protein, predicts a lipid anchor for membrane fixation. *The Journal of biological chemistry*, 269(21):14939–45, May 1994.

[McMeechan *et al.*, 2005] A McMeechan, M. A Lovell, T. A Cogan, K. L Marston, T. J Humphrey, and P. A Barrow. Glycogen production by different salmonella enterica serotypes: contribution of functional glgc to virulence, intestinal colonization and environmental survival. *Microbiology (Reading, Engl)*, 151(Pt 12):3969–77, Dec 2005.

[McNab, 2003] A McNab. Grid-based access control for unix environments, filesystems and web sites. *CiteSeerX - Scientific Literature Digital Library and Search Engine (United States)*, Preprint:24–28, Jan 2003.

[McNeil *et al.*, 2010] C McNeil, B Gallacher, C Harwood, J Hedley, P Manning, A Wipat, J Henderson, and N Keegan. Aptamems-id. http://gow.epsrc.ac.uk/ViewGrant.aspx?GrantRef=EP/G061394/1. *[accessed 27/07/2010]*, 2010.

[Medini *et al.*, 2008] D Medini, D Serruto, J Parkhill, D Relman, C Donati, R Moxon, and S. F. R Rappuoli. Microbiology in the post-genomic era. *Nature Reviews Microbiology*, 6:419–430, Jan 2008.

[Menne *et al.*, 2000] K. M Menne, H Hermjakob, and R Apweiler. A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics*, 16(8):741–742, Aug 2000.

[Mignot *et al.*, 2001] T Mignot, B Denis, E Couture-Tosi, A. B Kolstø, M Mock, and A Fouet. Distribution of s-layers on the surface of bacillus cereus strains: phylogenetic origin and ecological pressure. *Environmental Microbiology*, 3(8):493–501, Aug 2001.

[Miller *et al.*, 2009] M. E. B Miller, D. A Antonopoulos, M. T Rincon, M Band, A Bari, T Akraiko, A Hernandez, J Thimmapuram, B Henrissat, P. M Coutinho, I Borovok, S Jindou, R Lamed, H. J Flint, E. A Bayer, and B. A White. Diversity and strain specificity of plant cell wall degrading enzymes revealed by the draft genome of ruminococcus flavefaciens fd-1. *PLoS ONE*, 4(8):e6650, Jan 2009.

[Miron *et al.*, 2001] J Miron, D Ben-Ghedalia, and M Morrison. Invited review: adhesion mechanisms of rumen cellulolytic bacteria. *Journal of dairy science*, 84(6):1294–309, Jun 2001.

[Morrison and Field, 2010] N Morrison and D Field. http://www.environmentontology.org/. *[accessed 04/03/2010]*, 2010.

[Murphy *et al.*, 2007] K. M Murphy, P Travers, and M Walport. Jeneway's immunobiology. 2007.

[Nagano *et al.*, 2007] K Nagano, Y Murakami, K Nishikawa, J Sakakibara, K Shimozato, and F Yoshimura. Characterization of raga and ragb in porphyromonas gingivalis: study using gene-deletion mutants. *Journal of medical microbiology*, 56(Pt 11):1536–48, Nov 2007.

[Nagler-Anderson, 2001] C Nagler-Anderson. Man the barrier&excl; strategic defences in the intestinal mucosa. *Nature reviews immunology*, 1:59–67, Jan 2001.

[Nakai, 2000] K Nakai. Protein sorting signals and prediction of subcellular localization. *Advances in protein chemistry*, 54:277–344, Jan 2000.

[Nambu *et al.*, 1999] T Nambu, T Minamino, R. M Macnab, and K Kutsukake. Peptidoglycan-hydrolyzing activity of the flgj protein, essential for flagellar rod formation in salmonella typhimurium. *Journal of Bacteriology*, 181(5):1555–61, Mar 1999.

[Nataro *et al.*, 2005] J. P Nataro, P. S Cohen, H Mobley, and J. N Weiser. Colonization of mucosal surfaces. *ASM Press*, 2005.

[Natt *et al.*, 2004] N. K Natt, H Kaur, and G. P. S Raghava. Prediction of transmembrane regions of beta-barrel proteins using ann- and svm-based methods. *Proteins*, 56(1):11–18, Jul 2004.

[Navarre and Schneewind, 1999] W. W Navarre and O Schneewind. Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. *Microbiology and molecular biology reviews*, 63(1):174–229, Mar 1999.

[Neerincx and Leunissen, 2005] P. B. T Neerincx and J. A. M Leunissen. Evolution of web services in bioinformatics. *Briefings in Bioinformatics*, 6(2):178–188, Jun 2005.

[Neish, 2009] A Neish. Microbes in gastrointestinal health and disease. *Gastroenterology*, 136:65–80, Jan 2009.

[Nelson *et al.*, 2009] E. J Nelson, J. B Harris, J. G Morris, S. B Calderwood, and A Camilli. Cholera transmission: the host, pathogen and bacteriophage dynamic. *Nature Reviews Microbiology*, 7(10):693–702, Oct 2009.

[Newton and Snell, 1964] W Newton and E Snell. Catalytic properties of tryptophanase, a multifunctional pyridoxal phosphate enzyme. *Proceedings of the National Academy of Sciences*, 51:382–389, Mar 1964.

[Niemann, 2004] H Niemann. Adhesins and invasins of pathogenic bacteria: a structural view. *Microbes and Infection*, 6(1):101–112, Jan 2004.

[Noël *et al.*, 2010] C. J Noël, N Diaz, T Sicheritz-Ponten, L Safarikova, J Tachezy, P Tang, P.-L Fiori, and R. P Hirt. Trichomonas vaginalis vast bspa-like gene family: evidence for functional diversity from structural organisation and transcriptomics. *BMC Genomics*, 11:99, Jan 2010.

[Noppa *et al.*, 2001] L Noppa, Y Ostberg, M Lavrinovicha, and S Bergström. P13, an integral membrane protein of borrelia burgdorferi, is c-terminally processed and contains surface-exposed domains. *Infection and Immunity*, 69(5):3323–34, May 2001.

[O'Hara and Shanahan, 2006] A. M O'Hara and F Shanahan. The gut flora as a forgotten organ. *EMBO Reports*, 7(7):688–693, Jul 2006.

[Omaetxebarria *et al.*, 2007] M. J Omaetxebarria, F Elortza, E Rodríguez-Suárez, K Aloria, J. M Arizmendi, O. N Jensen, and R Matthiesen. Computational approach for identification and characterization of gpi-anchored peptides in proteomics experiments. *Proteomics*, 7(12):1951–1960, Jun 2007.

[O'Sullivan *et al.*, 2009] O O'Sullivan, J O'Callaghan, A Sangrador-Vegas, O McAuliffe, L Slattery, P Kaleta, M Callanan, G. F Fitzgerald, R. P Ross, and T Beresford. Comparative genomics of lactic acid bacteria reveals a niche-specific gene set. *BMC Microbiology*, 9:50, Jan 2009.

[Pajón *et al.*, 2006] R Pajón, D Yero, A Lage, A Llanes, and C. J Borroto. Computational identification of beta-barrel outer-membrane proteins in mycobacterium tuberculosis predicted proteomes as putative vaccine candidates. *Tuberculosis (Edinb)*, 86(3-4):290–302, Jan 2006.

[Pallen and Wren, 2007] M. J Pallen and B. W Wren. Bacterial pathogenomics. *Nature*, 449(7164):835–842, Oct 2007.

[Pallen *et al.*, 2003] M Pallen, R Chaudhuri, and I Henderson. Genomic analysis of secretion systems. *Current Opinion in Microbiology*, 6:519–527, Jan 2003.

[Park *et al.*, 2010] B. H Park, T. V Karpinets, M. H Syed, M. R Leuze, and E. C Uberbacher. Cazymes analysis toolkit (cat): web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using cazy database. *Glycobiology*, 20(12):1574–84, Dec 2010.

[Patti and Höök, 1994] J. M Patti and M Höök. Microbial adhesins recognizing extracellular matrix macromolecules. *Current opinion in cell biology*, 6(5):752–758, Oct 1994.

[Peterson and Mekalanos, 1988] K. M Peterson and J. J Mekalanos. Characterization of the vibrio cholerae toxr regulon: identification of novel genes involved in intestinal colonization. *Infection and Immunity*, 56(11):2822–9, Nov 1988.

[Peterson, 2002] K. M Peterson. Expression of vibrio cholerae virulence genes in response to environmental signals. *Current issues in intestinal microbiology*, 3(2):29–38, Sep 2002.

[Pierleoni *et al.*, 2006] A Pierleoni, P Martelli, P Fariselli, and R Casadio. Bacello: a balanced subcellular localization predictor. *Bioinformatics*, 22(14):e408, Jul 2006.

[Pohlschröder *et al.*, 2005a] M Pohlschröder, M Giménez, and K Jarrell. Protein transport in archaea: Sec and twin arginine translocation pathways. *Current Opinion in Microbiology*, 8:713–719, Jan 2005.

[Pohlschroder *et al.*, 2005b] M Pohlschroder, E Hartmann, N Hand, and K Dilks. Diversity and evolution of protein translocation. *Annual review of microbiology*, 18(7):91–111, Jan 2005.

[Ponnuraj *et al.*, 2003] K Ponnuraj, M. G Bowden, S Davis, S Gurusiddappa, D Moore, D Choe, Y Xu, M Hook, and S. V. L Narayana. A "dock, lock, and latch" structural model for a staphylococcal adhesin binding to fibrinogen. *Cell*, 115(2):217–28, Oct 2003.

[Postma *et al.*, 1993] P Postma, J Lengeler, and G Jacobson. Phosphoenolpyruvate: carbohydrate phosphotransferase systems of bacteria. *Microbiology Reviews*, 57(3):543–594, Jan 1993.

[Pridmore *et al.*, 2004] R. D Pridmore, B Berger, F Desiere, D Vilanova, C Barretto, A.-C Pittet, M.-C Zwahlen, M Rouvet, E Altermann, R Barrangou, B Mollet, A Mercenier, T Klaenhammer, F Arigoni, and M. A Schell. The genome sequence of the probiotic intestinal bacterium lactobacillus johnsonii ncc 533. *Proceedings of the National Academy of Sciences*, 101(8):2512–7, Feb 2004.

[Prost and Miller, 2008] L. R Prost and S. I Miller. The salmonellae phoq sensor: mechanisms of detection of phagosome signals. *Cellular microbiology*, 10(3):576–582, Mar 2008.

[Pruitt *et al.*, 2005] K Pruitt, T Tatusova, and D Maglott. Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 33(Database issue):D501–504, Jan 2005.

[Pruitt *et al.*, 2009] K. D Pruitt, T Tatusova, W Klimke, and D. R Maglott. Ncbi reference sequences: current status, policy and new initiatives. *Nucleic Acids Research*, 37(Database issue):D32–36, Jan 2009.

[Pull *et al.*, 2005] S. L Pull, J. M Doherty, J. C Mills, J. I Gordon, and T. S Stappenbeck. Activated macrophages are an adaptive element of the colonic epithelial progenitor niche necessary for regenerative responses to injury. *Proceedings of the National Academy of Sciences*, 102(1):99–104, Jan 2005.

[Qin *et al.*, 2006] X Qin, J. D Evans, K. A Aronstein, K. D Murray, and G. M Weinstock. Genome sequences of the honey bee pathogens paenibacillus larvae and ascosphaera apis. *Insect molecular biology*, 15(5):715–718, Oct 2006.

[Qin *et al.*, 2010] J Qin, R Li, J Raes, M Arumugam, K. S Burgdorf, C Manichanh, T Nielsen, N Pons, F Levenez, T Yamada, D. R Mende, J Li, J Xu, S Li, D Li, J Cao, B Wang, H Liang, H Zheng, Y Xie, J Tap, P Lepage, M Bertalan, J.-M Batto, T Hansen, D. L Paslier, A Linneberg, H. B Nielsen, E Pelletier, P Renault, T Sicheritz-Ponten, K Turner, H Zhu, C Yu, S Li, M Jian, Y Zhou, Y Li, X Zhang, S Li, N Qin, H Yang, J Wang, S Brunak, J Doré, F Guarner, K Kristiansen, O Pedersen, J Parkhill, J Weissenbach, M Consortium, M Antolin, F Artiguenave, H Blottiere, N Borruel, T Bruls, F Casellas, C Chervaux, A Cultrone, C Delorme, G Denariaz, R Dervyn, M Forte, C Friss, M v. d Guchte, E Guedon, F Haimet, A Jamet, C Juste, G Kaci, M Kleerebezem, J Knol, M Kristensen, S Layec, K. L Roux, M Leclerc, E Maguin, R. M Minardi, R Oozeer, M Rescigno, N Sanchez, S Tims, T Torrejon, E Varela, W d Vos, Y Winogradsky, E Zoetendal, P Bork, S. D Ehrlich, and J Wang. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, Mar 2010.

[Ragan, 2001] M Ragan. Detection of lateral gene transfer among microbial genomes. *Current Opinion in Genetics & Development*, 11:620–626, Jan 2001.

[Rahman *et al.*, 2008] O Rahman, S Cummings, D Harrington, and I. C Sutcliffe. Methods for the bioinformatic identification of bacterial lipoproteins encoded in the genomes of gram-positive bacteria. *World Journal of Microbiol Biotechnology*, 24:2377–82, Jan 2008.

[Rajilić-Stojanović *et al.*, 2007] M Rajilić-Stojanović, H Smidt, and W. M. D Vos. Diversity of the human gastrointestinal tract microbiota revisited. *Environmental Microbiology*, 9(9):2125–36, Sep 2007.

[Rashid *et al.*, 2007] M Rashid, S Saha, and G. P Raghava. Support vector machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. *BMC Bioinformatics*, 8(1):337–337, Sep 2007.

[Rasko *et al.*, 2005] D Rasko, M Altherr, C Han, and J Ravel. Genomics of the bacillus cereus group of organisms. *FEMS Microbiology Reviews*, 29:303–329, Jan 2005.

[Rasko *et al.*, 2008] D Rasko, M Rosovitz, and G Myers. The pangenome structure of escherichia coli: comparative genomic analysis of e. coli commensal and pathogenic isoloates. *Journal of Bacteriology*, 190(20):6881–93, Jan 2008.

[Rawlings *et al.*, 2008] N. D Rawlings, F. R Morton, C. Y Kok, J Kong, and A. J Barrett. Merops: the peptidase database. *Nucleic Acids Research*, 36(Database issue):D320–5, Jan 2008.

[Read *et al.*, 2003] T. D Read, G. S. A Myers, R. C Brunham, W. C Nelson, I. T Paulsen, J Heidelberg, E Holtzapple, H Khouri, N. B Federova, H. A Carty, L. A Umayam, D. H Haft, J Peterson, M. J Beanan, O White, S. L Salzberg, R c Hsia, G McClarty, R. G Rank, P. M Bavoil, and C. M Fraser. Genome sequence of chlamydophila caviae (chlamydia psittaci gpic): examining the role of niche-specific genes in the evolution of the chlamydiaceae. *Nucleic Acids Research*, 31(8):2134–47, Apr 2003.

[Ren and Paulsen, 2005] Q Ren and I. T Paulsen. Comparative analyses of fundamental differences in membrane transport capabilities in prokaryotes and eukaryotes. *PLoS Computational Biology*, 1(3):e27, Jan 2005.

[Rey *et al.*, 2005] S Rey, M Acab, J Gardy, M Laird, and K Defays. Psortdb: a protein subcellular localization database for bacteria. *Nucleic Acids Research*, 33(Database issue):D164–168, Jan 2005.

[Reyes *et al.*, 2010] A Reyes, M Haynes, N Hanson, F Angly, A Health, F Rohwer, and J Gordon. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*, 466:334–339, 2010.

[Riley *et al.*, 2007] M. L Riley, T Schmidt, I. I Artamonova, C Wagner, A Volz, K Heumann, H.-W Mewes, and D Frishman. Pedant genome database: 10 years online. *Nucleic Acids Research*, 35(Database issue):D354–7, Jan 2007.

[Rosenbach *et al.*, 2010] A Rosenbach, D Dignard, J. V Pierce, M Whiteway, and C. A Kumamoto. Adaptations of candida albicans for growth in the mammalian intestinal tract. *Eukaryotic Cell*, 9(7):1075–86, Jul 2010.

[Round and Mazmanian, 2009] J. L Round and S. K Mazmanian. The gut microbiota shapes intestinal immune responses during health and disease. *Nature reviews immunology*, 9(5):313–323, May 2009.

[Rubin *et al.*, 2007] D. L Rubin, N. F Noy, and M. A Musen. Protégé: A tool for managing and using terminology in radiology applications. *Journal of Digital Imaging*, 20(S1):34–46, Oct 2007.

[Rzhetsky *et al.*, 2004] A Rzhetsky, I Iossifov, T Koike, M Krauthammer, P Kra, M Morris, H Yu, P. A Duboue, W Weng, W. J Wilbur, V Hatzivassiloglou, and C Friedman. Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*, 37:43–53, Jan 2004.

[Rzhetsky *et al.*, 2008] A Rzhetsky, M Seringhaus, and M Gerstein. Seeking a new biology through text mining. *Cell*, 134(1):9–13, Jul 2008.

[Sainani, 2008] K Sainani. Mining biomedical literature: using computers to extract knowledge nuggets. *Biomedical computation review*, pages 17–27, Jun 2008.

[Saldanha, 2004] A Saldanha. Java treeview-extensible visualization of microarray data. *Bioinformatics*, 20(17):3246–8, Jan 2004.

[Saleh *et al.*, 2001] M Saleh, M Fillon, P Brennan, and J Belisle. Identification of putative exported/secreted proteins in prokaryotic proteomes. *Gene*, 269:195–204, Jan 2001.

[Salyers *et al.*, 2004] A Salyers, A Gupta, and Y Wang. Human intestinal bacteria as reservoirs for antibiotic resistance genes. *Trends in Microbiology*, 12(9):412–416, Jan 2004.

[Sampson and Gooday, 1998] M. N Sampson and G. W Gooday. Involvement of chitinases of bacillus thuringiensis during pathogenesis in insects. *Microbiology (Reading, Engl)*, 144 (Pt 8):2189–94, Aug 1998.

[Samudrala *et al.*, 2009] R Samudrala, F Heffron, and J. E McDermott. Accurate prediction of secreted substrates and identification of a conserved putative secretion signal for type iii secretion systems. *PLoS Pathogens*, 5(4):e1000375, Apr 2009.

[Sasaki *et al.*, 2002] Y Sasaki, J Ishikawa, A Yamashita, and K Oshima. The complete genomic sequence of mycoplasma penetrans, an intracellular bacterial pathogen in humans. *Nucleic Acids Research*, 30(23):5293–300, Jan 2002.

[Sayers *et al.*, 2010] E. W Sayers, T Barrett, D. A Benson, E Bolton, S. H Bryant, K Canese, V Chetvernin, D. M Church, M Dicuccio, S Federhen, M Feolo, L. Y Geer, W Helmberg, Y Kapustin, D Landsman, D. J Lipman, Z Lu, T. L Madden, T Madej, D. R Maglott, A Marchler-Bauer, V Miller, I Mizrachi, J Ostell, A Panchenko, K. D Pruitt, G. D Schuler, E Sequeira, S. T Sherry, M Shumway, K Sirotkin, D Slotta, A Souvorov, G Starchenko, T. A Tatusova, L Wagner, Y Wang, W. J Wilbur, E Yaschenko, and J Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 38(Database issue):D5–16, Jan 2010.

[Schäffer and Messner, 2005] C Schäffer and P Messner. The structure of secondary cell wall polymers: how gram-positive bacteria stick their cell walls together. *Microbiology (Reading, Engl)*, 151(Pt 3):643–651, Mar 2005.

[Schultz *et al.*, 2000] J Schultz, R. R Copley, T Doerks, C. P Ponting, and P Bork. Smart: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Research*, 28(1):231–234, Jan 2000.

[Schwebke and Burgess, 2004] J Schwebke and D Burgess. Trichomoniasis. *Clinical microbiology reviews*, 17(4):794–803, Jan 2004.

[Sekirov *et al.*, 2010] I Sekirov, S. L Russell, L. C. M Antunes, and B. B Finlay. Gut microbiota in health and disease. *Physiological reviews*, 90(3):859–904, Jul 2010.

[Seputiene *et al.*, 2003] V Seputiene, D Motiejūnas, K Suziedelis, H Tomenius, S Normark, O Melefors, and E Suziedeliene. Molecular characterization of the acid-inducible asr gene of escherichia coli and its role in acid stress response. *Journal of Bacteriology*, 185(8):2475–84, Apr 2003.

[Shah *et al.*, 2007] A Shah, V Markowitz, and C Oehmen. High-throughput computation of pairwise sequence similarities for multiple genome comparisons using scalablast. *Life Science Systems and and Applications Workshop*, pages 89–91, Jan 2007.

[Sharan *et al.*, 2007] R Sharan, I Ulitsky, and R Shamir. Network-based prediction of protein function. *Molecular Systems Biology*, 3:88, Jan 2007.

[Shoseyov *et al.*, 2006] O Shoseyov, Z Shani, and I Levy. Carbohydrate binding modules: biochemical properties and novel applications. *Microbiology and molecular biology reviews*, 70(2):283–295, Jun 2006.

[Slamti and Lereclus, 2002] L Slamti and D Lereclus. A cell-cell signaling peptide activates the plcr virulence regulon in bacteria of the bacillus cereus group. *The EMBO Journal*, 21(17):4550–9, Sep 2002.

[Slonim *et al.*, 2006] N Slonim, O Elemento, and S Tavazoie. Ab initio genotype–phenotype association reveals intrinsic modularity in genetic networks. *Molecular Systems Biology*, 2:1–14, Jan 2006.

[Smith and Nair, 2005] J Smith and R Nair. The architecture of virtual machines. *Computer*, pages 32–38, May 2005.

[Smith and Varzi, 2002] B Smith and A Varzi. Surrounding space the ontology of organism-environment relations. *Theory in Biosciences*, 120(2):139–162, Jan 2002.

[Smith *et al.*, 2007] B Smith, M Ashburner, C Rosse, J Bard, W Bug, W Ceusters, L. J Goldberg, K Eilbeck, A Ireland, C. J Mungall, N Leontis, P Rocca-Serra, A Ruttenberg, S.-A Sansone, R. H Scheuermann, N Shah, P. L Whetzel, and S Lewis. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–5, Nov 2007.

[Söding *et al.*, 2005] J Söding, A Biegert, and A. N Lupas. The hhpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, 33(Web Server issue):W244–8, Jul 2005.

[Soding, 2005] J Soding. Protein homology detection by hmm-hmm comparison. *Bioinformatics*, 21(7):951–960, Jan 2005.

[Song *et al.*, 2009] C Song, A Kumar, and M Saleh. Bioinformatic comparison of bacterial secretomes. *Genomics, proteomics and bioinformatics / Beijing Genomics Institute*, 7(1-2):37–46, Jun 2009.

[Sonnenburg *et al.*, 2005] J. L Sonnenburg, J Xu, D. D Leip, C.-H Chen, B. P Westover, J Weatherford, J. D Buhler, and J. I Gordon. Glycan foraging in vivo by an intestine-adapted bacterial symbiont. *Science*, 307(5717):1955–9, Mar 2005.

[Sonnenburg *et al.*, 2010] E Sonnenburg, H Zheng, P Joglekar, S. K Higginbottom, S. J Firbank, D. N Bolam, and J. L Sonnenburg. Specificity of polysaccharide use in intestinal bacteroides species determines diet-induced microbiota alterations. *Cell*, 141:1241–52, Jan 2010.

[Spinner *et al.*, 2008] J. L Spinner, J. A Cundiff, and S. D Kobayashi. Yersinia pestis type iii secretion system-dependent inhibition of human polymorphonuclear leukocyte function. *Infection and Immunity*, 76(8):3754–60, Aug 2008.

[Steinert *et al.*, 2000] M Steinert, U Hentschel, and J Hacker. Symbiosis and pathogenesis: Evolution of the microbe-host interaction. *Naturwissenschaften*, 87(1):1–11, Jan 2000.

[Stülke *et al.*, 1998] J Stülke, M Arnaud, G Rapoport, and I Martin-Verstraete. Prd–a protein domain involved in pts-dependent induction and carbon catabolite repression of catabolic operons in bacteria. *Molecular Microbiology*, 28(5):865–874, Jun 1998.

[Sun *et al.*, 2005] Y Sun, A Wipat, M Pocock, P Lee, and K Flanagan. Exploring microbial genome sequences to identify protein families on the grid. *TECHNICAL REPORT SERIES-UNIVERSITY OF NEWCASTLE UPON TYNE*, Jan 2005.

[Sutcliffe and Russell, 1995] I Sutcliffe and R Russell. Lipoproteins of gram-positive bacteria. *Journal of Bacteriology*, 177(5):1123–28, Jan 1995.

[Symersky *et al.*, 1997] J Symersky, J. M Patti, M Carson, K House-Pompeo, M Teale, D Moore, L Jin, A Schneider, L. J DeLucas, M Höök, and S. V Narayana. Structure of the collagen-binding domain from a staphylococcus aureus adhesin. *Nature structural biology*, 4(10):833–8, Oct 1997.

[Tabachnick and Fidell, 2001] B. G Tabachnick and L. S Fidell. Using multivariate statistics. *A Pearson Education Company*, 4th ed., 2001.

[Tatusov *et al.*, 2000] R. L Tatusov, M. Y Galperin, D. A Natale, and E. V Koonin. The cog database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28(1):33–36, Jan 2000.

[Tatusov *et al.*, 2003] R. L Tatusov, N. D Fedorova, J. D Jackson, A. R Jacobs, B Kiryutin, E. V Koonin, D. M Krylov, R Mazumder, S. L Mekhedov, A. N Nikolskaya, B. S Rao, S Smirnov, A. V Sverdlov, S Vasudevan, Y. I Wolf, J. J Yin, and D. A Natale. The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41–41, Sep 2003.

[Templeton, 2007] T Templeton. Whole-genome natural histories of apicomplexan surface proteins. *Trends in Parasitology*, 23(5):205–212, May 2007.

[Thomson *et al.*, 2006] N Thomson, S Howard, B Wren, and M Holden. The complete genome sequence and comparative genome analysis of the high pathogenicity yersinia enterocolitica strain 8081. *PLoS Genetics*, 2(12):2039–51, Jan 2006.

[Tjalsma *et al.*, 2000] H Tjalsma, A Bolhuis, J Jongbloed, and S Bron. Signal peptide-dependent protein transport in bacillus subtilis: a genome-based survey of the secretome. *Microbiology and Molecular Biology Reviews*, 64(3):515–547, Jan 2000.

[Tjalsma *et al.*, 2004] H Tjalsma, H Antelmann, J. D. H Jongbloed, P. G Braun, E Darmon, R Dorenbos, J.-Y. F Dubois, H Westers, G Zanen, W. J Quax, O. P Kuipers, S Bron, M Hecker, and J. M v Dijl. Proteomics of protein secretion by bacillus subtilis: separating the "secrets" of the secretome. *Microbiology and molecular biology reviews*, 68(2):207–233, Jun 2004.

[Todd *et al.*, 2001] A Todd, C Orengo, and J Thornton. Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology*, 307:1113–43, Jan 2001.

[Trost *et al.*, 2005] M Trost, D Wehmhöner, U Kärst, G Dieterich, J Wehland, and L Jänsch. Comparative proteome analysis of secretory proteins from pathogenic and nonpathogenic listeria species. *Proteomics*, 5(6):1544–57, Apr 2005.

[Tsuruoka *et al.*, 2008] Y Tsuruoka, J Tsujii, and S Ananiadou. Accelerating the annotation of sparse named entities by dynamic sentence selection. *BMC Bioinformatics*, 9 Suppl 11:S8, Jan 2008.

[Turnbaugh *et al.*, 2006] P. J Turnbaugh, R. E Ley, M. A Mahowald, V Magrini, E. R Mardis, and J. I Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027–31, Dec 2006.

[Turnbaugh *et al.*, 2007] P. J Turnbaugh, R. E Ley, M Hamady, C. M Fraser-Liggett, R Knight, and J. I Gordon. The human microbiome project. *Nature*, 449(7164):804–810, Oct 2007.

[Vadeboncoeur and Pelletier, 1997] C Vadeboncoeur and M Pelletier. The phosphoenolpyruvate:sugar phosphotransferase system of oral streptococci and its role in the control of sugar metabolism. *FEMS Microbiology Reviews*, 19(3):187–207, Feb 1997.

[van der Velden *et al.*, 1998] A. W v. d Velden, A. J Bäumler, R. M Tsolis, and F Heffron. Multiple fimbrial adhesins are required for full virulence of salmonella typhimurium in mice. *Infection and Immunity*, 66(6):2803–8, Jun 1998.

[van Haagen *et al.*, 2009] H. H. H. B. M v Haagen, P. A. C t Hoen, A. B Bovo, A d Morrée, E. M v Mulligen, C Chichester, J. A Kors, J. T d Dunnen, G.-J. B v Ommen, S. M v. d Maarel, V. M Kern, B Mons, and M. J Schuemie. Novel protein-protein interactions inferred from literature context. *PLoS ONE*, 4(11):e7894, Jan 2009.

[Vélez *et al.*, 2007] M. P Vélez, S. C. D Keersmaecker, and J Vanderleyden. Adherence factors of lactobacillus in the human gastrointestinal tract. *FEMS Microbiology Letters*, 276(2):140–148, Nov 2007.

[Vesanto *et al.*, 1996] E Vesanto, K Peltoniemi, T Purtsi, J. L Steele, and A Palva. Molecular characterization, over-expression and purification of a novel dipeptidase from lactobacillus helveticus. *Applied Microbiology Biotechnology*, 45:638–645, Jan 1996.

[von Mering *et al.*, 2007] C v Mering, P Hugenholtz, J Raes, S. G Tringe, T Doerks, L. J Jensen, N Ward, and P Bork. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, 315(5815):1126–30, Feb 2007.

[Wagner *et al.*, 1998] S Wagner, M. L Enss, M Cornberg, H Mix, S Schumann, G Kirchner, J Jähne, M. P Manns, and W Beil. Morphological and molecular characterization of human gastric mucous cells in long-term primary culture. *Pflugers Arch*, 436(6):871–881, Nov 1998.

[Walter *et al.*, 2009] M. C Walter, T Rattei, R Arnold, U Güldener, M Münsterkötter, K Nenova, G Kastenmüller, P Tischler, A Wölling, A Volz, N Pongratz, R Jost, H.-W Mewes, and D Frishman. Pedant covers all complete refseq genomes. *Nucleic Acids Research*, 37(Database issue):D408–11, Jan 2009.

[Wang and Granados, 1997] P Wang and R. R Granados. An intestinal mucin is the target substrate for a baculovirus enhancin. *Proceedings of the National Academy of Sciences*, 94(13):6977–82, Jun 1997.

[Wang *et al.*, 2003] X Wang, S. P Heazlewood, D. O Krause, and T. H. J Florin. Molecular characterization of the microbial species that colonize human ileal and colonic mucosa by using 16s rdna sequence analysis. *Journal of applied microbiology*, 95(3):508–520, Jan 2003.

[Weiser *et al.*, 2003] J. N Weiser, D Bae, C Fasching, R. W Scamurra, A. J Ratner, and E. N Janoff. Antibody-enhanced pneumococcal adherence requires iga1 protease. *Proceedings of the National Academy of Sciences*, 100(7):4215–20, Apr 2003.

[Wilson *et al.*, 2009] D Wilson, R Pethica, Y Zhou, C Talbot, C Vogel, M Madera, C Chothia, and J Gough. Superfamily–sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Research*, 37(Database issue):D380–6, Jan 2009.

[Wimley, 2002] W. C Wimley. Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. *Protein Science*, 11(2):301–312, Feb 2002.

[Wimley, 2003] W. C Wimley. The versatile beta-barrel membrane protein. *Current Opinion in Structural Biology*, 13(4):404–411, Aug 2003.

[Wipat *et al.*, 1996] A Wipat, N Carter, S. C Brignell, B. J Guy, K Piper, J Sanders, P. T Emmerson, and C. R Harwood. The dnab-phea (256 degrees-240 degrees) region of the bacillus subtilis chromosome containing genes responsible for stress responses, the utilization of plant cell walls and primary metabolism. *Microbiology (Reading, Engl)*, 142 (Pt 11):3067–78, Nov 1996.

[Wipat *et al.*, 2004] A Wipat, Y Sun, M Pocock, P Lee, P Watson, and K Flanagan. Developing grid-based systems for microbial genome comparisons: the microbase project. *Proceedings of the UK e-Science All Hands Meeting*, Jan 2004.

[Xu *et al.*, 2007] J Xu, M. A Mahowald, R. E Ley, C. A Lozupone, M Hamady, E. C Martens, B Henrissat, P. M Coutinho, P Minx, P Latreille, H Cordum, A. V Brunt, K Kim, R. S Fulton, L. A Fulton, S. W Clifton, R. K Wilson, R. D Knight, and J. I Gordon. Evolution of symbiotic bacteria in the distal human intestine. *PLoS Biology*, 5(7):1574–86, Jul 2007.

[Xu, 2010] D Xu. Cloud computing: An emerging technology. *Computer Design and Applications (ICCDA)*, 1:100–104, Jan 2010.

[Yu *et al.*, 2010] N. Y Yu, J. R Wagner, M. R Laird, G Melli, S Rey, R Lo, P Dao, S. C Sahinalp, M Ester, L. J Foster, and F. S. L Brinkman. Psortb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, 26(13):1608–15, Jul 2010.

[Zdobnov and Apweiler, 2001] E. M Zdobnov and R Apweiler. Interproscan–an integration platform for the signature-recognition methods in interpro. *Bioinformatics*, 17(9):847–848, Sep 2001.

[Zheng *et al.*, 2005] M Zheng, K Ginalski, L Rychlewski, and N. V Grishin. Protein domain of unknown function duf1023 is an alpha/beta hydrolase. *Proteins*, 59(1):1–6, Apr 2005.

[Zhou *et al.*, 2008] M Zhou, J Boekhorst, C Francke, and R Siezen. Locatep: Genome-scale subcellular-location predictor for bacterial proteins. *BMC Bioinformatics*, 9(1):173, Mar 2008.

[Zumaquero *et al.*, 2010] A Zumaquero, A. P Macho, J. S Rufián, and C. R Beuzón. Analysis of the role of the type iii effector inventory of pseudomonas syringae pv. phaseolicola 1448a in interaction with the plant. *Journal of Bacteriology*, 192(17):4474–88, Sep 2010.