# PROKARYOTIC PROTEIN SUBCELLULAR LOCALIZATION PREDICTION AND GENOME-SCALE COMPARATIVE ANALYSIS

by

Nancy Yiu-Lin Yu
B.Sc., University of British Columbia, 2003

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

In the
Department of Molecular Biology and Biochemistry
Faculty of Science

SIMON FRASER UNIVERSITY

Fall 2011

# Approval

**Name:** **Nancy Yiu-Lin Yu**

**Degree:** **Doctor of Philosophy**

**Title of Thesis:** **Prokaryotic protein subcellular localization prediction and genome-scale comparative analysis**

**Examining Committee:**

**Chair:** **Dr. Peter Unrau**
Associate Professor, Department of MBB

**Dr. Fiona S. L. Brinkman**
Senior Supervisor
Professor, Department of MBB

**Dr. Frederic Pio**
Supervisor
Associate Professor, Department of MBB

**Dr. Mark Paetzel**
Supervisor
Associate Professor, Department of MBB

**Dr. Lisa Craig**
Internal Examiner
Associate Professor, Department of MBB

**Dr. Michael Murphy**
External Examiner
Professor, Department of Microbiology and Immunology
University of British Columbia

**Date Defended/Approved:** **Wednesday, December 14, 2011**

# Abstract

Identifying protein subcellular localization (SCL) is important for deducing protein function, annotating newly sequenced genomes, and guiding experimental designs. Identification of cell surface-bound and secreted proteins from pathogenic bacteria may lead to the discovery of biomarkers, novel vaccine components and therapeutic targets. Characterizing such proteins for non-pathogenic bacteria and archaea can have industrial uses, or play a role in environmental detection. Previously, the Brinkman lab has developed PSORTb, the most precise SCL prediction software tool for bacteria. However, as we increasingly appreciate the diversity of prokaryotic species and their cellular structures, it became clear that there was a need to more accurately make predictions for more diverse microbes.

For my thesis research, I developed a new version of PSORTb that now provides SCL prediction capability for more prokaryotes, including Archaea and Bacteria with atypical cell wall and membrane structures. The new PSORTb also has significantly increased proteome prediction coverage for all bacterial species. The software is the first of its kind to predict subcategory localizations for bacterial organelles such as the flagellum as well as host cell destinations. Using both computational validations and a new proteomic dataset I produced, I established that PSORTb 3.0 outperforms all other published prokaryotic SCL prediction tools in terms of both precision and recall.

Furthermore, I have developed a semi-automated version of a comprehensive prokaryotic SCL database (PSORTdb) that provides access to experimentally verified and pre-computed SCL predictions for all sequenced prokaryotic genomes. I developed an ‘outer membrane prediction method’ which allows auto-detection of bacterial structure, distinguishing bacteria with one vs. two membranes. This method allows the database to be automatically updated as newly sequenced genomes are released. In addition, the method can aid more general analysis of a bacterial genome for which the bacteria’s associated cellular structure is not initially clear.

Finally, I performed a global analysis of SCL proportions for over 1000 sequenced bacterial and archaeal genomes. This is the most comprehensive SCL analysis of prokaryotes to date. My findings provide insights into prokaryotic protein network evolution, elucidate relationships between SCL proportions and genome size, and provide directions for future SCL prediction research.

**Keywords:** Bioinformatics; microbial genomics; proteomics; protein subcellular localization prediction; machine learning

# Acknowledgements

I would like to thank Dr. Fiona Brinkman who has provided guidance throughout the years, and all past and present members of the Brinkman laboratory, especially Matthew Laird, Matthew Whiteside, Erin Gill, William Hsiao, Morgan Langille, Raymond Lo, and Amber Feydnak, for insightful discussion and moral support during my graduate career. I would also like to thank my committee members, Dr. Mark Paetzel and Dr. Frederic Pio, for their encouraging words of support and helpful feedback for my thesis project. Thank you to all the collaborators of the PSORTb and PSORTdb project, including James Wagner, Phuong Dao, Cory Spencer, Dr. Sébastien Rey, Dr. Jennifer Gardy, Dr. Martin Ester, and Dr. Cenk Sahinalp.

Special thanks to Dr. Leonard Foster and members of the Foster lab for graciously hosting me and showing me state-of-the-art mass spectrometry experimental techniques as well as the associated computational analysis.

# Table of Contents

# List of Figures

# List of Tables

# Glossary

| | |
|---|---|
| **AA** | Amino acid |
| **Accuracy** | The proportion of true results (both positive and negative) in the population |
| **BLAST** | Basic local alignment search tool |
| **Coverage** | The proportion of a population or a genome that receives a prediction |
| **C** | Cytoplasmic |
| **CID** | Collision-induced dissociation |
| **CM** | Cytoplasmic membrane |
| **CSP** | Classically secreted protein |
| **CW** | Cell wall |
| **ESI** | Electrospray ionization |
| **EC** | Extracellular |
| **FN** | False negative |
| **FP** | False positive |
| **FT-ICR** | Fourier transform ion cyclotron resonance |
| **GO** | Gene Ontology |
| **HMM** | Hidden Markov model |
| **HPLC** | High performance liquid chromatography |
| **IT** | Ion trap |

| | |
|---|---|
| **KNN** | K-nearest neighbour, a classifying machine learning algorithm |
| **MALDI** | Matrix-assisted laser desorption ionization |
| **MCC** | Matthew’s correlation coefficient |
| **MS** | Mass spectrometry |
| **MS/MS** | Tandem mass spectrometry |
| **NN** | Neural network, a classifying machine learning algorithm |
| **NCSP** | Non-classically secreted protein |
| **NSP** | Non-secreted protein |
| **OM** | Outer membrane |
| **P** | Periplasm |
| **Q** | Quadrupole |
| **Precision/Specificity** | The proportion of true positives against all predicted positive results |
| **Recall/Sensitivity** | The proportion of true positives against all actual positive results |
| **SCL** | Subcellular localization |
| **SP** | Signal peptide |
| **SVM** | Support vector machine |
| **TAT** | Twin-arginine transporter |
| **TN** | True negative |
| **TP** | True positive |
| **TOF** | Time-of-flight |

# 1: Background

## 1.1 Bacterial cell structure

The domain of Bacteria consists of a phylogenetically and morphologically diverse group of single-celled organisms that are found in almost all types of habitats and environments on earth, from deep sea thermal vents, to ancient glaciers in Antarctica, to natural soils and waters, to radioactive wastes (Fredrickson *et al.*, 2004), as well as all over and within our bodies. Such broad diversity of bacteria is also reflected in their broad range of morphology. Although they appear in different shapes, colors and sizes, fundamentally they are structurally quite simple. Classically, a bacterial cell is typically enclosed in either one or two membranes. Bacteria with one membrane (monoderms) typically have a thick peptidoglycan cell wall encasing the membrane, giving it structure and some protection from the external environment. A second common bacterial structure consists of two layers of membranes (diderms), where the inner layer of cytoplasmic membrane is made of symmetrical lipid bilayer, and cell envelope or outer membrane is composed of asymmetrical lipid bilayer with lipopolysachharide (LPS). Between these inner and outer membranes is a gel-like space called the periplasm, where a thin layer of peptidoglycan (cell wall) resides. Cellular processes including certain types of metabolism, cell transport and cell defense and filtering mechanisms occur at this localization. Most bacteria discovered so far that can be cultured in the lab belong to one of these two groups, and can be distinguished by the classic Gram-staining technique.

### 1.1.1 Gram-stain technique for distinguishing monodermic and didermic bacteria

The Gram-stain technique, developed in 1884 by Hans Christian Gram (Beveridge, 2001), distinguishes Bacteria that can be cultured in the laboratory into two major structural groups: Gram-positive (which tends to detect monodermic bacteria with thick cell walls) and Gram-negative (which detects primarily didermic bacteria with thin cell walls). The staining works by applying a primary stain (normally crystal violet) to a heat-fixed bacterial culture of interest, adding a trapping agent (normally Gram's iodine), washing with acetone or alcohol, and counter staining the cells with safranin. Gram-positive bacteria typically possess a thick cell wall, which retains the crystal violet and appears purple under the microscope. Gram-negative bacteria usually possess a much thinner cell wall, which does not retain the crystal violet stain and so they appear pink due to the safranin counter stain. This technique has been widely accepted as the standard classification technique in microbiology and is still widely used today to differentiate the two major classic monodermic and didermic bacterial groups.

### 1.1.2 Bacteria with atypical Gram-positive and Gram-negative staining results

However, it is known that the Gram-stain procedure does not accurately identify all monodermic and didermic bacteria (Beveridge, 1990). Depending on the composition of the cell wall and the bacterium's growth phase, some bacteria are Gram-variable due to cell wall composition changes under different environmental conditions and/or life cycle. Furthermore, some bacterial groups do not fit into the classical Gram-positive or Gram-negative categories. For example, *Mycoplasma* spp. stain Gram-negative because they do not possess a peptidoglycan cell wall; however phylogenetically they are more related to the monoderms, which are predominantly Gram-positive (Miyata and Ogaki, 2006). In fact, bacteria that belong to the phylum Tenericutes (which includes *Mycoplasma* spp.) all have only one membrane with no murein cell wall enclosure. On the other hand, *Deinococcus* spp., which stains Gram-positive by the Gram stain technique, actually possesses an outer membrane (Thompson and Murray, 1981). This is because their cell walls are thicker than the ‘classical’ diderms, and the thick cell wall is able to retain the crystal violet dye. This issue will be revisited in chapter 2 and chapter 4 but in general, as we learn more about bacterial structural and phylogenetic diversity, the term “Gram-positive” and “Gram-negative” are not as useful in describing bacterial cellular structure. It is now proposed that “monoderm” and “diderm” more appropriately define the membrane structures of Bacteria (Sutcliffe, 2010).

### 1.1.3 Bacterial organelles

Although bacterial cells are not structurally as complicated as eukaryotes in terms of organelles, many bacteria do possess organelles that assist them in adapting to particular habitats. Common bacterial organelles include the flagella, which are found in diverse bacterial groups (Berg and Anderson, 1973). Most flagella in different bacterial organisms seem to have a common origin (Pallen *et al.*, 2005). Bacterial pili, on the other hand, have multiple origins. Some are used for adhesions; others are used in type IV secretion system for transferring DNA and proteins to other bacteria or to a eukaryotic host cell, discussed further in section 1.2.6. The bacterial type III secretion system apparatus is evolutionarily related to the bacterial flagellum, but has evolved into its own system specifically for protein translocation purposes (Ghosh, 2004). Bacteria from the phylum of Cyanobacteria often contain thylakoid membranes (Evans and Allen, 1973). Some Cyanobacteria species also contain gas vesicles that help them modulate buoyancy in water (Beard *et al.*, 1999). The phylum of Chloroflexi possesses chlorosomes, which are membraneous organelles with photosynthesis capabilities (Frigaard and Bryant, 2004). Bacteria from the phylum of Planctomycetes contain multiple membranous compartments (Fuerst and Sagulenko, 2011).

## 1.2 The bacterial protein sorting process

Proteins are structural and functional units within the cell, synthesized in the cytoplasm. In order to keep the bacterial cell alive with processes such as acquiring nutrients from the environment, sensing conditions in the environment, communicating with other bacteria, pumping out wastes and keeping harmful substances from entering the cell, bacteria have several secretion systems that target proteins to their destined cellular compartments or out into the extracellular milieu. This section describes the major secretion pathways within Bacteria – pathways that are key to ensuring proteins ends up in their correct subcellular locations.

### 1.2.1 General secretion pathway

The Sec secretion apparatus is found in all Bacteria, Archaea, and a homologous translocation system can be found in the endoplasmic reticulum membrane of eukaryotic cells. This is the major pathway that transports the majority of the proteins to or across the cytoplasmic membrane. The apparatus consists of SecA - an ATPase, SecYEG – the translocon, and SecM, which regulates SecA expression. There are two general routes in which a protein substrate is targeted to the SecYEG traslocon: an SRP-dependent pathway for co-translational protein translocation (Luirink *et al.*, 2005) and a SRP-independent pathway for post-translational protein translocation.

The translocation step works as follows: soluble chaperones, either SRP (for co-translational targeting), SecB (for post-translational targeting), or some other protein chaperone, recognize the signal sequence of protein substrates destined for export, and target the unfolded pre-protein to the Sec apparatus on the membrane (Fekkes and Driessen, 1999). Translocation occurs through the SecYEG translocon, which is a polypeptide conducting channel with a gating mechanism. A motor protein ATPase drives the translocation.

The signal peptide (SP) sequence, located at the N-terminus of proteins destined for the Sec-secretion pathway and recognized by the SRP (Lee and Bernstein, 2001), consists of a short, basic, positively-charged N-terminal region, a longer hydrophobic central region, and a short C-terminal hydrophilic region (von Heijne, 1990). The SP sequence is usually longer for non-cytoplasmic proteins in monoderms versus diderms (Nielsen *et al.*, 1997). For lipoproteins, the SPs are different. There is a cysteine residue at the cleavage site (von Heijne, 1990). SecB does not seem to recognize the SP sequence (Knoblauch *et al.*, 1999). Instead, it associates with a region of the substrate that is critical for protein folding (Bechtluft *et al.*, 2010).

### 1.2.2 Tat secretion pathway

Proteins transported by the Sec pathway must be unfolded in order to go through the SecYEG translocon (Driessen, 2001). There are some co-factor binding redox proteins that must be folded within the cytoplasm before they are transported across the cell (Weiner *et al.*, 1998; Saltikov and Newman, 2003). The Tat secretion pathway functions to deliver folded proteins across the cytoplasmic membrane in both monodermic and didermic bacteria. The Tat-apparatus has not been completely elucidated. However, it is known that it consists of the TatBC complex, which recognizes Tat signals and inserts the substrate into the membrane (Robinson and Bolhuis, 2004; Natale *et al.*, 2008). TatA mediates the translocation event, while a proton motif force serves as the energy source of the translocation machinery (Natale *et al.*, 2008).

The SP sequence for Tat substrates also consists of a tripartite structure, with a positively charged N-terminus, a longer hydrophobic region, and a C-terminal recognition sequence motif: S-R-R-X-F-L-K, where x denotes a polar amino acid. It tend to be longer than the signal peptide targeted for the Sec pathway, and the hydrophobicity of the central region of Tat SP tends to be lower than Sec SPs, due to having fewer glycines and threonines (Cristobal *et al.*, 1999). The C-terminus of the Tat SP contains basic residues, whereas the C-terminal regions of Sec SPs are almost never charged (Berks *et al.*, 2003).

### 1.2.3 The type I secretion system (T1SS)

The type I secretion system is a one-step substrate delivery system, where the substrate is transported from the cytoplasm directly to the extracellular environment. It consists of three major components: an ATP-binding cassette (ABC) transporter, a membrane fusion protein component, and outer membrane factor (OMF). The T1SS can be found in both monodermic and didermic bacteria (Franke *et al.*, 1999; Holland *et al.*, 2005).

The substrates of T1SS tend to be host colonization factors. Currently they are separated into two groups: RTX toxins and non-RTX toxins. RTX, which stands for repeats in toxins, and is found in S-layer proteins, metalloproteases, lipases, and poreforming cytotoxins (Heveker *et al.*, 1994; Stanley *et al.*, 1994). They share the common motif of GGXGXDXXX (where X could be any amino acid), which binds calcium ions (Letoffe and Wandersman, 1992). The secretion signals seem to be located at the C-terminus. They are specific but are non-conserved among the T1SS substrates (Holland *et al.*, 2005).

### 1.2.4 The type II secretion system (T2SS)

The type II secretion system secretes proteins in a 2-step process: it utilizes either the Sec pathway to transport unfolded substrates (Johnson *et al.*, 2006), or the Tat pathway to transport folded proteins to the periplasm, and then pseudopili pushes the substrate outside of the cell (Coulthurst and Palmer, 2008). A typical T2SS consists of 12-16 genes (Filloux, 2004), which includes the SecYEG translocon and a secretin pore in the outer membrane. T2SS is only found in Gram-negative Proteobacteria and is not a universal secretion system. The substrates of T2SSs consist of degradative enzymes and toxins. The T2SS is structurally similar to the Type IV secretion system. However the pseudopilus disassembles in T2SS but not in T4SS.

### 1.2.5 The type III secretion system (T3SS)

The type III secretion system (T3SS) is found only in Gram-negative bacterial species. So far 7 families of Type III secretion systems are known, commonly found to be encoded on genomic islands or plasmids (Troisfontaines and Cornelis, 2005). T3SS is a one-step delivery system that translocates substrates from the bacterial cell directly into the host cell. It consists of an injectisome, assembled by up to 25 proteins, 9 of which are conserved in all 7 families, and 8 of which are also found in the flagellum (Cornelis, 2006). First the injectisome itself is secreted to the outside of the cell, assembled into a needle structure, and then effectors are secreted in a particular order (Collazo and Galan, 1996; Wulff-Strobel *et al.*, 2002). Regulation of the secretion process happens at the transcriptional, post-transcriptional, translational levels, and also via conformational switching and secretion of the T3SS effectors (Aldridge and Hughes, 2002; Brutinel and Yahr, 2008; Deane *et al*., 2010). Substrate switch for the apparatus assembly appears to be affected by FliK/YscP as well as FlhB/YscU (Kutsukake *et al*., 1994; Williams *et al.*, 1996). There has been evidence showing that instead of being injected directly into the host cell, the effectors are seen as localized on the surface of the bacterial cell before being delivered into host cell via unknown mechanisms (Akopyan *et al.*, 2011). More studies are needed to elucidate this step.

The signals that direct substrates to the T3SS apparatus are still unclear. There have been evidence suggesting signals that direct effector proteins to the T3SS apparatus may occur at the mRNA level (Anderson and Schneewind, 1997; Ramamurthi and Schneewind, 2002). Other studies suggest that the signal sequence is located at the N-terminus of the effector proteins (Miao and Miller, 2000; Lloyd *et al.*, 2001). The N-terminus of the effectors do not seem to be conserved in terms of detectable homology; however, several computational analyses have been developed with the assumption that a signal sequence motif is located at the N-terminus (Arnold *et al.*, 2009; Samudrala *et al.*, 2009; Lower and Schneider, 2009; Wang *et al.*, 2011). Due to the complexity of a hierarchy in the order of substrate secretion for T3SS, the targeting process could be more complex than currently assumed, possibly involving chaperon proteins and accessory protein recognition, structural recognition, and other additional signals.

### 1.2.6 The type IV secretion system (T4SS)

The T4SS is a very versatile secretion system that is found both in monodermic and didermic bacteria, usually encoded on a plasmid. It is involved in DNA conjugation (DNA transfer from one bacterial cell to another), DNA uptake from and release into extracellular milieu, as well as protein effector delivery into targeted eukaryotic cells (Fullner *et al.*, 1996; Cascales and Christie, 2003; Alvarez-Martinez and Christie, 2009). Like the T1SS, the T4SS consists of three components, an energizing component, a core/channel apparatus, and a surface pilus component (Christie and Cascales, 2005; Christie, 2009). The T4SS is related to bacteriophage DNA injection machines, and the effector delivery system is often crucial for bacterial virulence. Phylogenetic classification of the type IV secretion system is challenging because many of the components appear to have gone through recombination and horizontal gene transfer in different species (Alvarez-Martinez and Christie, 2009). Classification by their respective functions is also difficult, because while some DNA conjugative systems only transfer DNA and some effector delivery systems specifically secrete proteins, certain T4SS have evolved to transfer both DNA and proteins. There has been no clear secretion signal sequences determined for T4SS substrates, but a non-truncated C-terminus appears to be important for T4SS substrate translocation (Nagai *et al.*, 2005; Vergunst *et al.*, 2005). However, for some species, the C-terminus contributes to the secretion process but is not essential for secretion by the T4SS (de Jong *et al.*, 2008).

### 1.2.7 The type V secretion system (T5SS)

The T5SS is considered as the simplest way to translocate a protein to the outside of the cell. Proteins that are secreted to the outer membrane surface and beyond are considered as part of T5SS if they 1) possess on their own all the information for translocation through cell envelope and/or require single accessory factors and 2) translocate across outer membrane via β-barrel proteins (Henderson *et al.*, 1998; Thanassi and Hultgren, 2000). It is divided into 3 classes: T5aSS, T5bSS, and T5cSS. The T5aSS is an auto-transporter, where the N-terminus consists of the substrate and the C-terminus forms a β-barrel pore in the outer membrane. The substrate, or passenger, passes through the β-barrel (or pores formed by β-barrel oligomers) and auto-cleaves itself (or is cleaved by an accessory protease located in the outer membrane) to be released into the extracellular milieu. The primary structure of proteins belonging to this category typically contains the following domains: a signal sequence at the N-terminus that allows it to be targeted across the cytoplasmic domain via the SecB-dependent Sec pathway, a functional motif (or passenger domain), which performs its functions once it is translocated across the outer membrane to the bacterial surface or cleaved to be released from the cell, an auto-chaperone domain, a linker region, and a C-terminal domain that folds into a β-barrel as it inserts into the outer membrane (Pohlner *et al.*, 1987; Desvaux *et al.*, 2004). It has been established that translocation of the passenger domain through the outer membrane requires the help of OMP85, an outer membrane biogenesis protein (Voulhoux and Tommassen, 2004)

The T5bSS is a two-partner transport system, where the passenger and translocator are two genes rather than one. The model system is the filamentous haemagglutinin (FHA) of *Bordetella pertussis* (Brown and Parker, 1987). The passenger gene and the translocator gene are denoted as TpsA and TpsB, respectively. TpsB has no homology with any of the T5aSS autotransporters, but it is part of the OMP85 family. TpsA and TpsB each have their own SPs targeting them to the Sec translocation pathway.

The T5cSS is a class of surface-attached oligomeric coiled coils adhesion (Oca) autotransporters (Roggenkamp *et al.*, 2003; Grosskinsky *et al.*, 2007). The C-terminus of proteins, or the putative transport domain in this group, consists of around 70 amino acids or four β-strands rather than the 14-15 β-strands that make up the C-terminal transport domain of T5aSS proteins. The linker region of this group has a coiled-coil structure, another feature that distinguishes this group from T5aSS (Grosskinsky *et al.*, 2007).

### 1.2.8 The type VI secretion system (T6SS)

T6SS was recently identified to be a new type of secretion system (Pukatzki *et al.*, 2006). It consists of 12 to >20 genes (Cascales, 2008), usually found on genomic islands. To date 5 distinct phylogenetic groups of T6SS has been found within didermic bacteria (Bingle *et al.*, 2008). The secretions system is a phage-related and is associated with virulence. Key components include an energizing unit – a protein with ATPase activity, an icmF homolog, a DotU (icmH) homolog, and ClpV. IcmF and DotU are also associated with the T4SS, however they seem to act as accessory proteins and are not essential to type IV secretion. Components of the apparatus that span the periplasm and the outer membrane are currently not known. The effectors seem to be delivered to eukaryotic cytoplasm (Rosales-Reyes *et al.*, 2011) as well as the periplasm of other bacteria (Russel *et al.*, 2011). The detailed mechanism of secretion is currently being actively studied by researchers.

### 1.2.9 The type VII secretion system (T7SS)

Also known as the ESX-1 secretion system (ESAT6-WXG100 secretion system), this was proposed by Abdallah *et al*. in 2007 to be re-named as the type VII secretion system (Abdallah *et al.*, 2007). So far it is found exclusively in *Mycobacterium* spp*.,* presumably to deal with translocating proteins across an outer membrane consisting of mycolic acids that is specific to the Mycobacterium genus. However, the secretion system has also been found in *Bacillus anthracis* (Garufi *et al.*, 2008) and *Staphylococcus aureus* (Burts *et al.*, 2005). The RD1 locus contains about 14 genes that are essential for the T7SS to function. The components include a cytoplasmic chaperone with an AAA+ ATPase domain, a subtilisin-like serine protease, several cytoplasmic membrane proteins, as well as proteins homologous to the FtsK/SpoIIIE family. The known effectors of this secretion system include ESAT-6 and CFP-10.

### 1.2.10 Some examples of other non-classical secretion systems

Besides the “classic” secretion systems with proteins spanning the inner and outer membranes, other types of protein secretion also exist. Diderms are known to secrete outer membrane vesicles (OMVs) containing packets of periplasmic material (Ellis and Kuehn, 2010). They are thought to be a mechanism to project metabolic function into the environment. OMV secretion is a widespread phenomenon in proteobacteria, but is mainly studied within pathogens like *Pseudomonas aeruginosa* (Bomberger *et al.*, 2009). These OMVs are projected into the environment via “nanopods”, which are bacterial surface structures that can “launch” these structures quite a distance away from the cell (Shetty *et al.*, 2011).

Recently a novel way of bacterial exchange of cytoplasmic molecules has been discovered (Dubey and Ben-Yehuda, 2011). Intracellular nanotubes have been observed to connect the cells of *Bacilus subtilis*, and even between species such as between *B. subtilis* and *Escherichia coli*. This previously uncharacterized bacterial structure can facilitate the delivery of DNA, metabolic molecules, and certain proteins from one bacterial cell to another.

### 1.2.11 Gram-positive (monodermic) secretion systems

Like the diderms, monoderms also use the general secretory (Sec) pathway to export most of its proteins out of the cytoplasm. It also uses Tat secretion system to export folded protein across the cytoplasmic membrane (Jongbloed *et al.*, 2002). It does not contain type III and type V secretion systems since these systems specifically deal with translocating proteins past the outer membrane. However, monoderms do possess ABC transporters not unlike the type I secretion systems and a pseudopilin export pathway similar to the type II and type IV secretion systems in diderms (Tjalsma *et al.*, 2000).

## 1.3 Archaeal cell structure, organelles and secretion systems

The cell structures and morphologies of the Archaea appear to be even more diverse than those in the domain of Bacteria. Fundamental structural differences between Archaea and Bacteria include membrane composition and cell wall composition (König, 1988). Archaeal membranes consist of phytanyl ether instead of fatty acid ester lipids, which are found in Bacteria as well as Eukarya. Unlike the bacterial peptidoglycan cell wall, the archaeal cell wall does not contain peptidoglycan. Most archaea possess a crystalline proteinaceous Surface layer (S-layer) (Kandler and Konig, 1978; Albers and Meyer, 2011). Only a few species identified to date lack an S-layer, but may possess a cell wall polymer made of pseudomurein, glycocalyx polysaccharide, glutaminylglycan, sulphated heteropolysaccharide, halomucin, and other variations not found in Bacteria (Albers and Meyer, 2011). So far, only one archeal genus, namely *Ignicoccus* spp., has been found to possess two layers of membranes (and no S-layer) (Huber *et al.*, 2000) . As more archaeal species are discovered, more cell envelope structural diversity may emerge.

### 1.3.1 Archaeal organelles

Like Bacteria, several archaeal groups possess a flagellum. However, the flagellar components and assembly machinery resemble bacterial type IV pili more than the bacterial flagellar apparatus (Ng *et al.*, 2006). The main function of the archaeal flagellum is for motility purposes. Other types of type IV-related pili include ones that mediate cellular aggregation when induced by UV light (Frols *et al.*, 2008), and a bindosome that is involved in carbohydrate uptake (Zolghadr *et al.*, 2011). Only one type of archaeal pilus has been found so far that does not belong to the type IV pilus group. It is a pilus found in *Methanothermobacter thermoautotrophicus* that is around 5 nm in diameter (Thoma *et al.*, 2008). It seems to mediate surface adhesion and cell-cell contacts in biofilm.

Besides the flagellum and pili, other archaeal structures that have been studied so far include the cannulae (Horn *et al.*, 1999) and the hami (Moissl *et al.*, 2005). The cannulae are hollow tubes 25 nm in diameter and as long as 30-150 µm. They seem to stem from the quasi-periplasmic space of *Pyrodicitium* spp. and are not connected to the cell cytoplasm (Nickell *et al.*, 2003). Their precise functions are unknown but are thought to be involved in cell-cell attachment and anchoring. Hami are filamentous cell appendages found on euryarchaeon SM1 cells (Moissl *et al.*, 2005). They are around 7-8 nm in diameter, with hook-like structures along the appendages. Like the cannulae, they are also thought to be used for attachment to other cells to create archaeal communities.

### 1.3.2 Archaeal secretion systems

The general Sec secretion pathway is found in Archaea as well. The archaeal Sec translocator appears to contain both bacterial and eukaryotic components (Pohlschroder *et al.*, 1997). The Tat secretion pathway is also found in Archaea and secretes both C-terminally anchored lipoproteins as well as soluble proteins (Gimenez *et al.*, 2007). Some archaeal species, for example some *Thermococcus spp.*, some *Sulfolobus spp.*, and *Aciduliprofundum boonei* are known to secrete membrane vesicles into the medium (Soler *et al.*, 2008; Ellen *et al.*, 2009; Reysenbach *et al.*, 2006).

### 1.3.3 Side note: are Archaea really prokaryotes?

The world of cellular organisms has been divided into two categories: organisms without a membrane-bound nucleus called “prokaryotes”, and organisms containing a membrane-bound nucleus called “eukaryotes” (Stanier and van Niel, 1962). The concept was that all bacteria came from the same prokaryotic ancestor. In 1977, Carl Woese’s phylogenetic analysis paper placed the Archaea in a separate domain from Bacteria and Eukarya (Woese and Fox, 1977). Further phylogenetic analysis has revealed that, although organisms in the domain Archaea do not contain a membrane-bound nucleus like the Bacteria, phylogenetically they do not share closer ancestry to Bacteria than to

Eukarya. A few papers have argued that the term “prokaryote” is obsolete, since current data supports a three branch evolution of life rather than eukaryote/prokaryote dichotomy (Woese, 2004; Pace, 2006). In terms of evolution and phylogeny, it does not make sense to put Archaea and Bacteria into one group. This thesis deals primarily with organism structural and subcellular compartments, from a cellular structure perspective, it makes sense to apply the structural definition of “prokaryote” to Archaea, which has the same cellular compartment organizations as monodermic bacteria. I do not wish to imply though that this infers a phylogenetic relationship between these well-defined Domains of life.

## 1.4 The importance of SCL identification and prediction

One of the goals of basic microbiology research is to figure out the functions of all genes within a bacterial genome, so we can better understand the organism. By determining the SCL of a protein, we can narrow down its potential set of functions. From a more practical point of view, SCL identification and prediction has many useful purposes. To combat pathogenic bacteria, researchers are interested in identifying drug targets that are bacterial surface proteins (Solis and Cordwell, 2011; Vogel and Claus, 2011), as well as vaccine components that are cell surface associated or secreted, since these are more accessible to the immune system. Given the genome of a bacterium can be easily sequenced and its protein-encoding genes relatively accurately identified (versus for eukaryotes), it is desirable to identify all the microbe’s cell surface proteins, to help create a shortlist of potential drug targets and vaccine components (Relman, 2011). In addition, surface and secreted proteins can also serve as diagnostic biomarkers. SCL prediction practicality does not only apply to pathogens. For environmental bacteria, identifying cell surface and secreted proteins from whole-genome as well as metagenomics data can be used for environmental detection biomarkers (Wu *et al.,* 2009). Some of these secreted proteins are often enzymes that can be adapted for useful industrial applications (Niehaus *et al.*, 1999). For example, certain enzymes from Archaea are heat and acid stable and are useful in laundry detergents (de Champdore *et al.,* 2007) and in pretreatment of biomass in the production of biofuel (Graham *et al.*, 2011). There is interest in discovering more of these enzymes that could have important commercial use.

Within microbiology research, SCL prediction is also helpful for researchers, who may be working with a gene of unknown function and would like to know its destined SCL to facilitate protein extraction or general experimental design and confirmation. SCL prediction is now also routinely used as part of a microbial genome annotation pipeline. There is a growing trend in proteomics projects as well, where subcellular fractions of bacteria are submitted to high-throughput identification (Walther and Mann, 2010). Given there are always possible contaminations and experimental errors (Rey *et al.*, 2005b), high precision computational SCL prediction can serve as a valuable means for evaluation and confirmation purposes. In the next two sections, laboratory experimental and computational means to identify/predict SCLs will be discussed.

## 1.5 Examples of laboratory experimental approaches for determining protein SCL

A protein's SCL has been traditionally determined in the laboratory via techniques suitable for low throughput, single protein analysis only. However, more recently a number of high-throughput identification methods have become available. A list of common experimental approaches will be briefly reviewed in the following section.

### 1.5.1 Microscopy-based visualization techniques

Protein SCL may be identified by labeled antibodies. The protein of interest is detected by primary antibodies. Secondary antibodies, labeled by colloidal gold for electron microscopy or for fluorescent microscopy, labeled by green fluorescent protein (GFP) that binds to the primary antibodies, allow researchers to visually identify the protein's localization (Chalfie *et al.*, 1994). A limitation to this method is that only proteins with available antibodies for specific tagging can be detected this way. Also, sometimes with this visualizing method, the identification of SCL can still be ambiguous, especially for peripheral proteins. Because electron microscopy is done on thin slices of a cell, it may not accurately represent living cell's conditions. For example, slide preparation may cause shifts of peripheral membrane proteins that are not tightly bound to the membrane, or it may cause cytoplasmic proteins to gather at the inner membrane surface.

### 1.5.2 PhoA fusion technique

Alkaline phosphatase PhoA fusion can be used to screen for proteins that are exported to the cytoplasmic membrane and beyond. PhoA, an enzyme which is folded and active in the periplasm can be fused in-frame with a gene of interest into a plasmid and expressed in *E. coli*. A positive result of a PhoA activity assay would indicate that the gene of interest is indeed an exported protein. This method has a small false positive rate, and has been shown to successfully evaluate about 310 of *Pseudomonas aeruginosa* protein-encoded genes (Lewenza *et al.*, 2005). However, not all exported proteins can be identified using this method. Some PhoA fusion proteins may be toxic to the bacterial cell, perhaps due to folding problems or the protein is not meant to be expressed at the level as dictated by the promoter used in this method.

### 1.5.3 Subcellular fractionation and two-dimensional (2D) gel electrophoresis

The 2D-gel electrophoresis is a common technique that has been used to identify proteins from a given subcellular fraction (Lescuyer *et al.*, 2004). It is usually coupled with mass spectrometry protein identification, as described in the next section. However, more recently there is a trend in moving away from 2D gel electrophoresis and moving towards other types of sub-fractionation (separating complex samples of proteins), since 2D gel protein identification has lower sensitivity and requires larger amounts of sample than methods in which proteins are separated in solution. Regardless, cellular fractionation-based approaches can have problems with contamination of abundant proteins in fractions that the abundant protein is not actually associated with (for example, certain abundant cytoplasmic proteins contaminating a secreted fraction due to cell lysis).

### 1.5.4 Mass spectrometry identification of subcellular fractions

Mass spectrometry (MS) is an analytical technique that can accurately identify a molecule by measuring it mass-to-charge ratio (m/z). It has been in use for over 100 years, but has become feasible for analyzing peptides and protein molecules after the maturation of the development of “soft ionization” techniques by Fenn *et al*. in 1989 (Fenn *et al.*, 1989). MS-based proteomics has started to replace 2-D gel electrophoresis because of its higher accuracy, resolution and throughput (Walther and Mann, 2010). There are a large variety of MS machines, but in general, a mass spectrometer is comprised of three main components: an ion source, a mass analyzer, and a detector. The different types of technologies used in each component are listed in Table 1.1.

Table 1.1 Types of technologies for each principle components of a mass spectrometer

| **Ion Source** | **Mass Analyzer** | **Detector** |
|---|---|---|
| ESI | Quadrupole (Q) | Electron Multiplier |
| MALDI | Quadrupole Ion Trap (IT) | Array Detector |
|  | Linear Ion Trap (LIT) |  |
|  | Time of Flight (TOF) |  |
|  | Tandem Time of Flight (TOF/TOF) |  |
|  | Fourier Transform Ion Cyclotron Resonance (FT-ICR) |  |
|  | Orbitrap |  |

There are two types of "soft-ionization" techniques used for proteomics-based mass spectrometry: electrospray ionization (ESI) (Manisali *et al.*, 2006; Mann and Fenn, 1992) and matrix-assisted laser desorption ionization (MALDI) (Hillenkamp *et al.,* 1991). ESI is usually carried out in conjunction with nanospray high-performance liquid chromatography (HPLC). Essentially, both methods produce ionized proteins or peptide molecules that are then analyzed by the mass analyzer. To identify the protein or peptide, usually tandem mass-spec is used (MS/MS). The most commonly used peptide fragment technique is collision-induced dissociation (CID) (Griffiths *et al.*, 2001). The first mass-spec identifies the mass-to-charge (m/z) ratio of a whole protein or a peptide, and the second mass-spec further fragments the polypeptide preferentially at the amide bond between the amino acids. The mass difference between the peptides one amino acid apart can be used to identify the amino acids in the peptide sequence.

There are many types of mass analyzers, one of the simplest available being the Time of Flight (TOF) instruments (Dorosenko *et al*, 1999). The ions are accelerated by an electric field into a long vacuum tube. The time it takes for the ions to travel across this fixed-length tube is proportional to its m/z ratio. It tends to have lower resolution than other types of mass analyzers. The quadrupole (Q) instrument is comprised of four metal rods that are equidistance apart from each other, with a quadrupolar electric field running through the rods (Mann *et al*., 2001). Only ions of a certain m/z ratio can pass through to the detector. This machine can be coupled with an ion trap, or alternatively, a triple quadrupole (Q1, Q2, Q3) can perform tandem MS: Q1 selects for ions of a certain m/z to go through; Q2 acts as a collision cell to fragment the ionized peptide, and Q3 separates and allows the fragmented peptides to be resolved and thus identify the amino acid sequence (Graham *et al.*, 2011). The ion trap (IT) analyzer consists of a rotating three-dimensional electrostatic field. As opposed to Q analyzers that sequentially allow ions of a certain m/z to get through the instrument, all ionized peptides are initially trapped in the IT. The electric field is manipulated such that only ions of a certain m/z are ejected from the trapping field and move towards the detector (Domon and Aebersold, 2006). The Fourier transform ion cyclotron resonance (FT-ICR) instrument is the most expensive, but has the highest resolving power for samples as low as attamoles ($10^{-18}$ mol) (Patrie *et al*., 2004). Ions are injected into a cell perpendicular to a static magnetic field. Ions in the field assume a cyclotron motion in a frequency correlated to their m/z. A Fourier transform is then carried out to calculate the m/z of the ionized peptides. The orbitrap is a newer invention, developed in the late 1990s by Makarov (Hu *et al.*, 2005). The trap consists of a split outer electrode and an inner electrode. Ions injected into the orbitrap orbit the central electrode while oscillating in the z direction. The frequency of the oscillations produces an image current on the split outer electrode, which can be resolved into m/z by Fourier transformation. The orbitrap is highly accurate and has high resolution like the FT-ICR, but is cheaper and therefore is a common instrument used in current proteomic applications.

Typically, a protein sample to be analyzed by MS/MS is cleaved into peptides using trypsin, which is known to cleave proteins after lysine residues. For analysis of proteins from organisms with known genomes/predicted proteome, once the machine has identified the amino acids within a peptide, the mass spectra from that peptide is searched against a library of theoretical spectra of all peptides generated from a theoretical trypsin cleavage deduced from the organism's deduced proteome. Any molecular labeling of the protein and possible post-translational modification of the proteins are also taken into account in this library. To estimate the rate of error from matches to the wrong peptide mass spectrum by chance, all the measured spectra are also searched against an error estimation library. This library is constructed by reversing the amino acid sequence order of the deduced proteome trypsin-cleaved peptide library. The number of hits to this error-checking library provides a good estimate of the false-positive identification rate of the MS/MS analysis (Graham *et al.,* 2011).

## 1.6 Computational approaches for predicting bacterial and archaeal protein SCL

### 1.6.1 A brief history of computational prediction of protein SCL

The first bacterial protein SCL prediction software that made predictions for multiple localizations was published by Kenta Nakai and Minoru Kanehisa in 1991 (Nakai and Kanehisa, 1991). It consisted of an expert system which employed a collection of "if-then" rules, based on a list of protein sequence features that provided clues to a protein's probable SCL, including features like sorting signals, hydrophobicity, sequence motifs, etc. If there were competing rules that applied to a protein, only one rule would be selected using a "conflict resolution" procedure. This system predicted four SCLs in Gram-negative bacteria: cytoplasmic, cytoplasmic membrane, periplasm, and outer membrane. The accuracy of this method was found to be 65%. Since then, other SCL prediction software has been developed (discussed below), but because most software forced a prediction result for all input proteins, the accuracy remained relatively low. For example, Nakai and Kanehisa's PSORT I method did not at all predict secreted proteins, and so secreted proteins would, 100% of the time, be miss-predicted as another localization due to its forced predictions as one of the method's four localizations. For certain purposes such as accurately identifying a set of outer membrane proteins as potential drug or vaccine targets, it would be desirable to have a software tool that can generate SCL predictions with higher confidence. Thus in 2003, Gardy *et al.* associated with the Brinkman lab developed PSORTb 1.0, a novel protein SCL predictor for Gram-negative bacteria (Gardy *et al.*, 2003). Like PSORT I, PSORTb uses a number of modules of algorithms that can predict one or multiple SCLs with high specificity. These modules include SCL-BLAST for motif analysis, PROSITE for motif-based analysis, HMMTOP for transmembrane α-helix prediction, an outer membrane motif analysis, an SP predictor to differentiate cytoplasmic vs. non-cytoplasmic proteins, and a variation of SubLoc, a support vector machine (SVM)-based algorithm. More details about these algorithms are described in the following section. The results of each module are entered into a Bayesian network to produce the final prediction score. The Bayesian network serves to give weight to the prediction value of each module based on the module's measured accuracy. The usage of a Bayesian network greatly boosted the software's overall precision in comparison to PSORT I's decision tree-based method. In 2005, PSORTb version 2 was extended to predict both Gram-negative and Gram-positive bacteria SCLs with over 95% precision (Gardy *et al.*, 2005). It contained an expanded training dataset, as well as newly implemented SVMs using protein subsequences instead amino acid composition as classifying features. This version predicts five SCLs for didermic bacteria: cytoplasmic, cytoplasmic membrane, periplasm, outer membrane, and extracellular. For monodermic bacteria, the software predicts four SCLs: cytoplasmic, cytoplasmic membrane, cell wall, and extracellular. The improvements to PSORTb 2.0 significantly increased its recall and prediction coverage compared to PSORTb 1.0. An explosion of SCL predictors have been developed since then. However, most of such software has focused on maximizing accuracy, providing predictions for all input proteins (i.e. forcing predictions) at the expense of precision, making it less useful for biologists wanting reliable prediction results. The only other SCL prediction software like PSORTb that emphasized precision and did not force predictions (i.e. would return a result of "unknown" if the method could not precisely deduce the protein SCL) was Proteome Analyst (PA) (Lu *et al.*, 2004). PA combines a homology and a text-mining approach to predict bacterial SCLs. Given an input protein sequence, it first compares the sequence against the Swiss-Prot database using BLAST, identifies a set of annotated homologs, and then uses a Naïve-Bayes classifier to predict the SCL based on the Swiss-Prot annotations. The algorithm first determines the probability that each Swiss-Prot annotated functional keyword is associated with an SCL. These computed probabilities are then used in predicting the SCL of an input protein. The Naïve-Bayes algorithm will be discussed in the next section. Proteome Analyst also has multiple versions, where the newest version (PA 3.0) outperforms PSORTb 2.0. However, it is unpublished and the improvements made to the new version remain unknown.

### 1.6.2 Computational algorithms for bacterial SCL prediction

All computational SCL predictions use some sort of algorithm to classify and generate predictions given input protein(s) of interest. This section discusses common algorithms that have been used for SCL prediction. A list of recently published bacterial SCL prediction software is shown in Table 1.2.

Table 1.2 List of published bacterial SCL prediction software and methods. See also http://www.psort.org for a larger list of methods that may be used to aid SCL predictions.

| **Software** | **Predicted SCLs for diderms** | **Predicted SCLs for monoderms** | **Analytical method** | **Accepts multiple input sequences?** | **Availability** |
|---|---|---|---|---|---|
| PSORT I (Nakai and Kanehisa, 1991) | C, CM,P,OM | C, CM, EC | Multi-component | No | Web server |
| PSORTb 3.0 (Yu *et al.*, 2010) | C, CM,P,OM, EC,( Flagellum, Fimbrium, T3SS, Host, Spore) | C, CM, CW, EC | Multi-component | Yes | Standalone software and web server |
| Proteome Analyst (Lu *et al.*, 2004) | C, CM,P,OM, EC | C, CM, EC | BLAST+Naïve Bayes | Yes | Web server not available |
| NClassG+ (Restrepo-Montoya *et al.,* 2011) | N/A | Non-EC, EC | Multi-component | No | Web server not working |
| SignalP 4.0 (Petersen *et al.*, 2011) | C, not C | C, not C | NN | No | Web server |
| SecretP 2.0 (Yu, L. *et al.,* 2010) | CSP, NCSP, NSP* | CSP, NCSP, NSP* | Multi-component | No | Web server not working |
| SecretomeP 2.0 (Bendtsen *et al.*, 2004a) | Non-EC, EC | Non-EC, EC | Neural network | Yes | Web server not working |
| Gpos-mPLoc (Shen and Chou, 2009) | N/A | C, CM, CW, EC | Multi-component | No | Web server |
| Gneg-mPLoc (Shen and Chou, 2010) | C, CM,P, OM, EC, Flagellum, Fimbrium, nucleoid | N/A | Multi-component | No | Web server |
| SubcellPredict (Niu *et al.*, 2008) | C,P, EC | C,P, EC | AdaBoost | N/A | Web server not available |

| | | | | | |
|---|---|---|---|---|---|
| PSLDoc (Chang *et al.*, 2008) | C, CM,P, OM, EC | C, CM, CW, EC | SVM | Yes | Standalone software |
| PSL101 (Su *et al.*, 2007) | C, CM,P,OM, EC | N/A | SVM | Yes | Web server |
| SOSUI-GramN (Imai *et al.*, 2008) | C, CM,P,OM, EC | N/A | Linear discriminant analysis | Yes | Web server |
| SLP-Local (Matsuda *et al.*, 2005) | C,P, EC | C,P, EC | | Yes | Web server |
| SubLoc (Hua and Sun, 2001) | C,P, EC | C,P, EC | SVM | Yes | Web server not available |
| CELLO 2.5 (Yu *et al.*, 2006) | C, CM,P,OM, EC | C, CM, CW, EC | SVM | Yes | Web server |
| PSLpred (Bhasin *et al.,* 2005) | C, CM,P,OM, EC | None | SVM | No | Web server |
| LOCtree (Nair and Rost, 2005) | C,P, EC | C,P, EC | SVM | Yes | Web server |
| P-CLASSIFIER (Wang *et al.,* 2005) | C, CM,P,OM, EC | None | SVM | Yes | Web server not available |
| FFT-based SCL predictor (Wang *et al.,* 2007) | C,P, EC | None | Fourier transform SVM | N/A | No software available |
| GNBSL (Guo *et al.*, 2006) | C, CM,P,OM, EC | None | SVM | N/A | Web server not available |
| HensBC (Bulashevska and Eils, 2006) | C, CM,P,OM, EC | None | Bayesian classifier + Markov chain models | N/A | No software available |

*CSP = classically secreted protein; NSCP = non-classically secreted protein; NSP = non-secreted protein

#### 1.6.2.1 Decision Tree

Nakai and Kanehisa's first attempt at bacterial SCL prediction used a decision tree algorithm, which uses a tree structure (Nakai and Kanehisa, 1991). Each node is a decision node (some sort of "If-then" test) or a leaf node (an SCL). Starting at the root of the tree, the input data (protein sequence) gets classified at each decision node until it reaches a SCL prediction. While the algorithm is informative and easy to interpret by humans, it is not quite sophisticated enough for accurately classifying SCLs from protein sequences, as (for example) it does not weigh each decision branch point.

#### 1.6.2.2 Hidden Markov model (HMM) algorithm

The hidden Markov model (HMM) algorithm is an extension of weighted matrices, where sequence alignments used to calculate the weights allowed for features such as gaps, which means motifs of various lengths can be taken into consideration in the calculations. HMM has been used extensively for sequence alignment and motif identification/profile generation purposes, and is used for SP identification (Nielsen and Krogh, 1998; Bendtsen *et al.*, 2004b) and the prediction of α-helices of cytoplasmic membrane proteins (Tusnady and Simon, 2001; Krogh *et al.*, 2001; Kall *et al.*, 2004; Viklund and Elofsson, 2004).

#### 1.6.2.3 K-Nearest neighbor (KNN) algorithm

The KNN algorithm is a classification method that assigns classification based on the majority vote of the object's neighbors, majority being the label of k-nearest neighbors. A problem with KNN is that it is sensitive with the local structure of data. That is, if the training dataset is not evenly distributed, new cases may be misclassified.

Shen and Chou created Gpos-PLoc (Shen and Chou, 2007) and Gneg-PLoc (Chou and Shen, 2006) using optimized evidence theoretic KNN (OET-KNN) by hybridizing GO (gene ontology) terms with amphiphilic pseudo amino acid composition (PseAA) algorithms (Chou, 2005). In 2009 and 2010, Shen and Chou extended their software to allow for multiple localization predictions (Shen and Chou, 2009; Shen and Chou, 2010).

#### 1.6.2.4 Neural network (NN) algorithm

The neural network (NN) algorithm is inspired by biological networks of neurons, where input data are presented to layers of artificial "neurons", which compute weighted sums of the inputs and then a nonlinear function is applied to the sum. Unlike weighted matrices, the calculation of the scores can be nonlinear. This allows correlation between positions in the sequence to influence the final prediction. Reinhardt and Hubbard were the first to employ the NN algorithm on the prediction of protein SCLs for both prokaryotes and eukaryotes (Reinhardt and Hubbard, 1998). NN was also used to predict signal peptides that direct a protein to the cell membrane and beyond. The original SignalP for signal peptide prediction was created using NN algorithm (Nielsen *et al.*, 1997). The second version employed an HMM algorithm (Nielsen and Krogh, 1998); for SignalP 3.0 the performance was improved by combining the prediction results of the HMM and the NN (Bendtsen *et al.*, 2004b), and very recently, a fourth version of SignalP was published using a pure NN algorithm that out-performs all existing signal peptide prediction methods (Petersen *et al.*, 2011).

#### 1.6.2.5 Alignment-based prediction

It has been shown that sequence-identity acts as a highly accurate predictor for SCL prediction (Nair and Rost, 2002; Yu *et al.*, 2006). A 25% sequence identity produces around 70% precision while >30% sequence identity provides 80-90%+ SCL prediction precision, which is higher than most machine-learning based SCL prediction algorithms (Yu *et al.*, 2006). The draw-back, of course, is that this method cannot be applied to proteins with little or no sequence similarity to proteins with known SCLs.

#### 1.6.2.6 Support vector machine (SVMs)-based methods

The support vector machine algorithm is a binary classification method. Take a set of training examples, where each item (in our case a sequence) is labeled as x or y (in our case x and y are SCLs). The SVM algorithm then builds a model based on the features of the training data. This model is constructed as mapped points in multi-dimensional space such that when test data in entered, it is assigned to the same map to see which side of the classification it falls upon. This has been found to be very useful in determining localization, probably because multiple features are usually associated with dictating a given localization. This method has been widely tested by many studies, some of which are mentioned here.

Hua and Sun first used SVMs to build a bacterial SCL predictor called SubLoc (Hua and Sun, 2001), using amino acid compositions as the classifying feature. SVMs have since gained popularity among SCL prediction researchers. PSORTb 1.0 initially used a version of SubLoc in its predictions, and had since then developed a frequent subsequence-based SVM predictor for PSORTb 2.0 and PSORTb 3.0 (She *et al.*, 2003). Other software that uses SVMs to predict SCL includes the following:

SLP-Local (Matsuda *et al.*, 2005) splits an input protein sequence into N-terminal, middle, and C-terminal sections, and uses a combination of single amino acid (AA), twin AAs, and local frequencies of distance between successive basic, hydrophobic, and other AAs to build the SVM models. It does not make predictions though for inner or outer membrane proteins.

PSLpred (Bhasin *et al.*, 2005) uses PSI-BLAST and three SVMs based on AA composition, di-peptide composition, and physicochemical properties to predict SCLs for Gram-negative bacteria.

P-Classifier (Wang *et al.*, 2005) uses n-peptide and physicochemical properties as sequence features to build SVMs.

GNBSL (Gram-negative bacterial subcellular localization) (Guo *et al.*, 2006) builds a position-specific frequency matrix (PSFM) and a position-specific scoring matrix (PSSM) for each input protein sequence by searching against the Swiss-Prot database. Whole sequence AA composition, N- and C- terminus AA composition, dipeptide composition, and segment composition are used as sequence features. A combination of probabilistic neural network (PNN) and SVMs are constructed to build this SCL predictor.

CELLO 2.0 (Yu *et al.*, 2006) uses a two-level SVM to predict SCLs. In the first level, for each SCL, Yu *et al.* constructed SVMs based on n-peptide composition, partitioned AA composition, n-gapped dipeptide composition, and local AA composition. Each SVM generates a probability distribution. The second level SVM acts as a jury in determining the final probability of each SCL. The final version of CELLO combines the SVMs and a homology-based method to make predictions.

PSL101 (Su *et al.*, 2007) uses a hybrid of one-versus-one SVM models and secondary structural homology to predict SCLs. To build SVM models, a positive training dataset and a negative training dataset is needed for each SVM. Most software employ a one-versus-rest approach, where the positive dataset is a set of protein sequences of that known SCL, and the negative dataset comprises of proteins from all other SCLs. This paper used AA composition, dipeptide composition, relative solvent accessibility, and secondary structure element encoding schemes to build 1-vs-1 SCL SVMs.

PSLDoc (Chang *et al.*, 2008) uses gapped-dipeptides as sequence features for creating the SVM classifiers, and combines this with probabilistic latent semantic analysis (PLSA) to predict SCLs.

NClassG+ (Restrepo-Montoya *et al.*, 2011) uses frequencies, dipeptides, physicochemical factors and PSSM as sequence transformation vectors (sequence features) to build SVM classifiers for non-classically secreted Gram-positive bacterial proteins.

#### 1.6.2.7 Naïve Bayes classifier

The naïve Bayes classifier is a simple probabilistic classifier that applies the Baye's theorem (conditional probability) with the naïve assumption of independence between each individual classifying feature. In spite of its over-simplified design, this method seems to work quite well, even if the individual properties used to make classifications are dependent on real world situations. As noted above, the most precise methods available, PSORTb and PA, both utilize a naïve Bayes classifier to generate a final prediction. HensBC (hierarchical ensembles of Bayesian classifiers) (Bulashevska and Eils, 2006) is another example that predicts SCLs by constructing Bayesian classifiers based on Markov chains built from amino acids. However because its software is not released, we cannot test it against PSORTb and PA.

#### 1.6.2.8 Other algorithms

Wang *et al.* employed fast Fourier transform (FFT) on three types of amino acid representation models: the c-p-v matrix (composition, polarity, molecular volume), the EIIP model (electron-ion interaction potential, hydrophobicity, and stability of α-helix formation), and a hybrid of the two models (Wang *et al.*, 2007), and used these three models as sequence features for training SVMs. The authors claim that FFT is suitable for processing biological signals and show that the prediction results are somewhat comparable to older methods such as SubLoc.

AdaBoost is a type of boosting strategy that combines weak predictors to improve prediction results. SubcellPredict (Niu *et al.*, 2008) uses Random Forests as a weak classifier and AdaBoost to boost its performance value. Its performance was compared against basic NN and SVM algorithms but not with any other published software methods.

SOSUI-GramN (Imai *et al.*, 2008) uses physicochemical properties of N- and C-terminal regions of proteins, as well as whole proteins to predict SCLs. It consists of an inner membrane predictor and a number of modules that distinguishes the non-cytoplasmic membrane proteins into other SCLs.

### 1.6.3 Performance comparison between existing bacterial SCL prediction software

All researchers that use SCL prediction software are curious about performance comparisons between various SCL prediction software tools. There are many problems in conducting such a comparison for all software methods. For one thing, not all software makes predictions for all SCLs for monodermic and didermic bacteria. Some software tools choose to only provide predictions for soluble protein SCLs (cytoplasmic, periplasmic, extracellular), presumably for the reason that the algorithm development for cytoplasmic membrane proteins are already quite mature and precise. Other software only accepts single protein sequences as input, making it very difficult to run predictions on a batch set of protein data. To be able to compare different software fairly, ideally a testing dataset of bacterial protein sequences is required that are of low similarity to each other and are not part of the training dataset of any of the software methods being compared. This dataset is not easy to come by. Normally it requires a person to manually review biological literature to find proteins with experimentally-verified, previously unknown SCLs, or generate new laboratory experimental data of high quality regarding protein SCLs for proteins of previously unknown localization.

Performance metrics that commonly appear in computational SCL prediction papers include the following:

**Precision/Specificity**, or true positive rate, is defined as the rate of making a prediction correctly. Mathematically, it is described as TP/(TP+FP), where TP stands for true positive and FP stands for false positive.

**Recall/Sensitivity** is defined as the rate of making a relevant prediction out of all the test instances. The emphasis here is on the *number* of instances that are identified.

Mathematically, it is described as TP/(TP + FN), where TP stands for true positive and FN stands for false negative.

**Accuracy** is a balance between precision and recall, or (TP + TN) / (TP + TN + FP + FN), where TP stand for true positive, FP stands for false positive, TN stands for true negative, and FN stands for false negative.

**Matthew's correlation coefficient (MCC)** is a measure of the quality of binary classifications. It takes both false positives and false negatives into account.

$$MCC = \frac{TP \bullet TN - FP \bullet FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

**Coverage** is defined as the proportion of prediction made over all the proteins within the protein-coding genome of an organism. The difference between recall and coverage is that recall is measured over a test set of proteins, carefully chosen to not overlap with the training dataset and to be non-redundant, and it examines correct predictions, whereas coverage is measured over a biologically-relevant dataset such as a whole genome, and is only useful if the precision of the method being analyzed is known and the method does not force predictions.

Precision and recall (or specificity and sensitivity) often oppose each other. To make a very precise predictor, one often needs to sacrifice recall, and to maximize recall, or even to maximize accuracy often the false positive rate increases, reducing precision. Most SCL prediction software developed by computer scientists tends to focus on maximizing accuracy and MCC, whereas biologists tend to not be comfortable using software with high false positive rates. Only a few SCL prediction software tools, namely

PSORTb and PA, are designed to focus on maintaining a certain high level of precision, while increasing overall accuracy/recall in successive versions. Because of the difference in design intentions, comparisons within the research papers obviously each focus on their design strengths.

In order to calculate precision, recall, accuracy and MCC, different methods can be used. Commonly, a k-fold cross-validation or a leave-one-out cross-validation (LOOCV) is used to test the performance of a prediction algorithm. K-fold cross validation involves partitioning the training dataset into k subsets. One of the subsets is used to test the algorithm, while the remaining k-1 subsets are combined and used to build the machine learning algorithm model. This is done k times with each of the subsets. In LOOCV, each sample, in this case protein sequence is used as test data, while the rest of the data are used to train the algorithm. This is repeated until every sequence in the training dataset has been independently tested. This is the same as a k-fold cross-validation, with k equal to the size of the entire training dataset. LOOCV is a more exhaustive test than a 5-fold or 10-fold cross validation, but is very computationally time consuming and not practical in all situations.

Rey *et al.* showed that SCL prediction methods are now exceeding the accuracy of certain high-throughput laboratory proteomic methods for SCL prediction (Rey *et al.*, 2005b). Gardy and Brinkman showed in a 2006 review of methods that PSORTb 2.0 has a precision higher than all other bacterial SCL prediction software of its time using an independent set of 299 proteins that should not be in the training dataset of any of the prediction software (Gardy and Brinkman, 2006). In Chapter 2, the above mentioned metrics are used to compare PSORTb with other comparable bacterial protein SCL prediction software published since 2006.

### 1.6.4 Archaeal SCL prediction methods

Before the start of this thesis project, there was no available SCL prediction software specifically for Archaea. Proteome Analyst contains pre-computed SCL results for two archaeal species, *Methanobacterium thermoautotrophicum* and *Methanococcus jannaschii* (http://webdocs.cs.ualberta.ca/~bioinfo/PA/GOSUB/archaea.html), but currently does not provide a running SCL prediction server. A signal peptide predictor has been made more recently for Archaea which performs better than using prediction tools trained on bacterial or eukaryotic data (Bagos *et al.*, 2009).

## 1.7 Prokaryotic protein SCL datasets and databases

A searchable database of experimentally verified and/or computationally predicted SCLs would be very useful for researchers, both SCL prediction algorithm developers and molecular biologists. The dataset of experimentally verified protein SCLs would serve as training and testing dataset for prediction software, and the pre-computed SCL results can serve to annotate genome sequences, to verify experimental results, and as a comparison with high-throughput proteomics experiments. The SignalP group published its training dataset (Nielsen *et al.*, 1997), which has been used by many SCL prediction research groups for building and testing their software. The first database of annotated SCLs for over 60,000 prokaryotic and eukaryotic was DBSubLoc (Guo *et al.*, 2004). However their website does not seem to be available any more. PSORTdb was created in 2005 containing both a set of proteins with known SCL as well as pre-computed SCL predictions of PSORTb 2.0 for sequenced genomes published by NCBI (Rey *et al.*, 2005a) and was manually updated a couple of times a year as new genomes became available. Two databases specific for Gram-positive pre-computed protein SCLs, Augur (Billion *et al.*, 2006) and LocateP (Zhou *et al.*, 2008), were published in 2006 and 2008, respectively. Augur contains surface protein predictions only, and is presented as a pipeline where user can retrieve a list of Gram-positive surface proteins with specific types of domains and sorting signals for a given Gram-positive bacterial species. LocateP also specifically predicts different types of cell surface proteins. However it also provides prediction results for cytoplasmic and extracellular proteins. CobaltDB (Goudenege *et al.*, 2010) is a client-server application for download that integrates protein SCL prediction results from a number of prokaryotic SCL predictors, signal peptide (SP) predictors and SP cleavage site predictors. It provides a set of pre-computed results for a number of sequenced prokaryotic genomes, and does not appear to have been updated since its publication.

## 1.8 Proposal for improving subcellular localization prediction and global comparative analysis

Though many bacterial protein SCL prediction methods have been developed, none are as precise or as user-friendly as PSORTb, in that both a web-based server and standalone version are available; pre-computed prediction results are available; single protein or batch submissions of proteins for analysis are possible, and predictions are not forced and are usually correct when made due to the high precision. This has been considered as a method-to-beat in terms of generating SCL prediction improvements. But none of the existing SCL software compared in 2006 (Gardy and Brinkman, 2006) took into account bacterial organelles or bacteria with atypical cell wall and membrane structures, including PSORTb. Furthermore, no SCL prediction software existed for archaeal proteins. For my thesis project, I therefore set out to improve PSORTb by (A) adding capabilities for detecting proteins targeted to more specialized subcellular compartments/organelles for more diverse bacteria, (B) adding archaeal protein prediction capabilities for the first time, (C) improving overall prediction sensitivity and proteome coverage of PSORTb while maintaining high (>95%) precision (the latter requiring me to generate a new dataset of proteins of known localization through a laboratory proteomics method). I also aimed to (D) build a database of the most comprehensive experimentally verified SCLs as well as pre-computed SCLs for sequenced prokaryotic genomes as they become available through NCBI, the first database of its kind that would be continually updated in a more automated fashion. Finally, I wished to (E) perform a global analysis of SCL proportion trends for a broad group of sequenced prokaryotic genomes, with the hypothesis that through such an analysis I may better understand the fundamental protein network structure of Bacteria and Archaea. Chapter 2 presents a published paper that describes the results of objectives (A), (B) and (C). Chapter 3 presents a published paper for objective (D), and Chapter 4 presents the results of objective (E). Chapter 5 discusses the implications of these results and the future of prokaryotic SCL prediction/identification research.

# 2: PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes

**Published in:** Bioinformatics, 2010, 26(13):1608-1615.

**Author List:** Nancy Y. Yu, James R. Wagner, Matthew R. Laird, Gabor Melli, Sébastien Rey, Raymond Lo, Phuong Dao, S. Cenk Sahinalp, Martin Ester, Leonard J. Foster and Fiona S. L. Brinkman

**Author contributions**: NYY and SR performed literature search to expand the training dataset. NYY updated the code for the new version of PSORTb, performed parts of the mass spectrometry experiment, and wrote the manuscript. JRW optimized the new SVMs. MRL put together the new PSORTb software package. GM contributed to expanding the training dataset. RL prepared the subcellular fraction protein samples for proteomics analysis. PD wrote the protein feature extraction program for the SVMs. LJF generated the final dataset using mass spectrometry. FSLB contributed to the design of the new software, the analyses and co-wrote the paper.

## 2.1 Abstract

**Motivation:** PSORTb has remained the most precise bacterial protein subcellular localization (SCL) predictor since it was first made available in 2003. However, the recall needs to be improved and no accurate SCL predictors yet make predictions for Archaea, nor differentiate important localization subcategories, such as proteins targeted to a host cell or bacterial hyperstructures/organelles. Such improvements should preferably be encompassed in a freely available web-based predictor that can also be used as a standalone program.

**Results:** We developed PSORTb version 3.0 with improved recall, higher proteome-scale prediction coverage, and new refined localization subcategories. It is the first SCL predictor specifically geared for all prokaryotes, including Archaea and bacteria with atypical membrane/cell wall topologies. It features an improved standalone program, with a new batch results delivery system complementing its web interface. We evaluated the most accurate SCL predictors using 5-fold cross validation plus we performed an independent proteomics analysis, showing that PSORTb 3.0 is the most accurate but can benefit from being complemented by Proteome Analyst predictions.

**Availability:** http://www.psort.org/psortb (download open source software or use the web interface).

**Contact**: psort-mail@sfu.ca.

**Supplementary Information:** Supplementary data are available at *Bioinformatics* online and in Appendix 1.

## 2.2 Introduction

Computational prediction of bacterial protein subcellular localization (SCL) provides a quick and inexpensive means for gaining insight into protein function, verifying experimental results, annotating newly sequenced bacterial genomes, detecting potential cell surface/secreted drug targets, as well as identifying biomarkers for microbes. In recent years, this area of computational research has achieved an impressive level of precision (Gardy and Brinkman, 2006), allowing SCL prediction tools to be reliably integrated into automated proteome annotation pipelines and to complement analyses of high-throughput proteomics experiments.

PSORTb version 2.0 (Gardy *et al.*, 2005), the most precise bacterial SCL prediction software (Gardy and Brinkman, 2006), was introduced in 2005, and has been widely used for the SCL prediction of individual proteins as well as for whole proteomes. It generates prediction results for five major localizations for Gram-negative bacteria (cytoplasmic, inner membrane, periplasmic, outer membrane, extracellular) and four localizations for Gram-positive bacteria (cytoplasmic, cytoplasmic membrane, cell wall, extracellular). Since then, numerous SCL prediction tools have been created for bacteria using a variety of machine learning algorithms: CELLO version 2.0 (Yu *et al.*, 2006) uses multi-layered Support vector machines (SVMs); SLP-Local predicts SCLs based on local composition and distance frequencies of amino acid groups (Matsuda *et al.*, 2005); PSL101 makes predictions based on amino acid compositions coupled with structural feature conservations (Su *et al.*, 2007), and PSLDoc bases its SVM features on gapped di-peptides (Chang *et al.*, 2008). Other tools such as Gpos-PLoc (Shen and Chou, 2007) and Gneg-PLoc (Chou and Shen, 2006) make predictions for bacterial proteins by clustering Swiss-Prot proteins with annotated SCLs based on their Gene Ontology (GO) terms and amino acid properties using the K-nearest neighbor (KNN) algorithm. Some methods, such as SubcellPredict and HensBC, combine multiple classifying algorithms in order to boost the prediction performance (Niu *et al.*, 2008; Bulashevska and Eils, 2006). LocateP (Zhou *et al.*, 2008) and Augur (Billion *et al.*, 2006) differentiate between different types of membrane-anchored, cell wall anchored and secreted proteins for Gram-positive bacterial proteomes. Based on the principle that training datasets could benefit from being genus specific, TBPred (Rashid *et al.*, 2007) was developed specifically for the genus of *Mycobacterium spp*.

Even though many bacterial SCL prediction methods have been published, most of them focus on optimizing prediction accuracy - maximizing the number of positive predictions on the training dataset, at the expense of producing more false positive results. Furthermore, none of the current bacterial SCL predictors provide standalone versions of software for users. Most web servers also do not provide convenient means for analyzing whole bacterial proteomes. PSORTb remains one of the most user-friendly bacterial SCL prediction tools, providing both a web server and a standalone version, and allowing for both single and batch sequence processing. Its accompanying database, PSORTdb (Rey *et al.,* 2005), provides a dataset of experimentally verified protein localizations, as well as pre-computed prediction results for more than 1000 sequenced bacterial genomes available from NCBI. Because of its focus on maintaining high precision, it does not return a forced prediction if the localization score does not reach a minimum cut-off. As a result, only about 50% of proteins encoded in Gram-negative bacterial genomes and about 75% of proteins encoded in Gram-positive bacterial genomes receive a prediction from PSORTb. Thus, there is a need to produce an updated version with better genome coverage.

The current localization classifications for PSORTb and most existing SCL prediction software do not provide any information on proteins targeted to specialized bacterial hyperstructures/organelles such as the flagellum, the fimbrium/pilus, or proteins destined to the host cell. Gneg-PLoc (Chou and Shen, 2006) attempts to address this by providing prediction categories for the nucleosome (DNA-binding proteins) and the flagellum. Gpos-PLoc (Shen and Chou, 2007) provides predictions for Gram-positive periplasmic proteins. Some studies have attempted to predict effector proteins secreted by the type III secretion system based on N-terminal signal sequence of proteins (Arnold *et al.,* 2009; Samudrala *et al.,* 2009). Ideally, comprehensive SCL prediction software should incorporate predictions for these more specialized compartments in addition to reporting major SCLs.

Typically, bacterial organisms that stain Gram-positive consist of one cytoplasmic membrane and a thick cell wall, whereas a Gram-negative organism is enclosed by a thin cell wall within a periplasm and an outer membrane that surrounds the entire cell. However, some bacteria have cell structures that do not fit with the classical Gram-negative or Gram-positive cell model. For example, *Mycoplasma* spp. and other members of the phylum Tenericutes stain Gram-negative, yet they have no outer membrane or cell wall (Miyata and Ogaki, 2006). *Deinococcus* spp. has a thick cell wall and is considered as a Gram-positive organism, but they also have an outer membrane (Thompson and Murray, 1981). Therefore, to make protein subcellular localization predictions for all prokaryotes, not only does an archaeal predictor need to be created, but we also need to be able to make a predictor that can handle the four possible bacterial cell structures that we now know are possible: Gram-positive without an outer membrane (i.e. traditional Gram-positives), Gram-negative with an outer membrane (i.e. traditional Gram-negatives), Gram-positive with an outer membrane, and Gram-negative without an outer membrane. Only then is a predictor able to cover the true diversity of prokaryotic life, which will become more important as increased sampling of prokaryotes occurs through metagenomics and other projects (Wu *et al.,* 2009).

In addition to bacterial SCL prediction algorithms, several software packages for predicting SCL of eukaryotic proteins have been developed, despite the fact that they are much harder to predict due to the greater complexity of eukaryotic cells (see http://www.psort.org for a list of available eukaryotic protein SCL predictors). However, there are no dedicated SCL prediction tools for Archaea, the third domain of life whose basic cellular compartments are similar to that of a Gram-positive bacterium. Not only do they represent an entire domain of abundant organisms that inhabit the earth, they produce many thermotolerant and halotolerant enzymes that have wide industrial applications (de Champdore *et al.,* 2007). Furthermore, identification of novel cell surface and secreted proteins can also be very helpful for designing new methods for the detection of specific archaeal species in the environment.

To address these issues, we have created PSORTb version 3.0, with a significant increase in recall of predictions as well as proteome prediction coverage while maintaining >95% precision (see *2.3 Software evaluations – using literature and Swiss-Prot-based datasets* for definitions of precision and recall). In addition, we recognize that the current localization classification scheme does not adequately cover all bacterial proteins' detailed localization sites. Therefore, we have added new localization subcategories commonly found in many groups of bacteria – the first subcategory localization system for an SCL predictor. Options specifically for predicting archaeal proteins and proteins in organisms with membrane structures not reflecting Gram stains have also been implemented. We further improved usability by adding an online batch submission system with formatted results returned by email. For the standalone version, we have simplified the installation procedure. Finally, we examined the results of combining complementary SCL predictions in order to produce accurate predictions for the majority of prokaryotic proteomes, using an independent, proteomics-derived laboratory test dataset to aid the analysis.

## 2.3 Methods

### 2.3.1 Training Dataset

The training dataset contains data from ePSORTdb 2.0 (Rey *et al.,* 2005), which was used to build PSORTb 2.0, Swiss-Prot version 49 (Wu *et al.*, 2006), plus protein localization data obtained from manual literature search (the latter comprises 30% of the dataset). From Swiss-Prot, protein localizations were based on the 'Comments – Subcellular location' field with review. A natural language processing predictive model, TeGRR (Melli *et al.*, 2007), was used as a text mining technique on literature abstracts to confirm the validity of the Swissprot SCL annotation. Organisms were separated into Gram-positive and Gram-negative groups based on their phylum/class and literature review. Bacteria belonging to the phyla of Actinobacteria, Chloroflexi, Deinococcus-Thermus of the order Thermales, Firmicutes of class Bacilli and most Clostridia were categorized as Gram-positive bacteria. Bacteria in phylum groups not mentioned above were categorized as Gram-negative. For proteins from the Swissprot library with annotated subcellular locations, those labeled as ‘fragment’, ‘by similarity’, ‘probable’, and ‘potential’ were removed. Those that were annotated with very specialized localizations such as ‘chlorosome’ and ‘chromatophore’ were not used for this dataset. Proteins that were labeled with ambiguous terms such as ‘cell envelope’ were manually confirmed for their specific localization if possible, or discarded if the precise localization could not be determined. Some protein entries were manually retrieved from the literature as well as the EcoSal database (http://www.ecosal.org) and the *Pseudomonas* Genome Database (Winsor *et al.*, 2008). The archaeal testing dataset was obtained in a similar fashion as the bacterial dataset. The training dataset for building the archaeal predictor was created by combining archaeal proteins with Gram-positive/Gram-negative cytoplasmic and cytoplasmic membrane proteins, as well as Gram-positive cell wall and extracellular proteins, as this was found to notably increase accuracy when evaluated using archaeal proteins. In total, the Gram-negative training dataset has expanded from 1572 proteins to 8230 proteins; the Gram-positive dataset has increased from 576 to 2652 proteins, and 810 archaeal proteins have been added to the training dataset. The full training dataset is available at http://www.psort.org/dataset/datasetv3.html.

### 2.3.2 Software implementation and updates

#### 2.3.2.1 New localization subcategories

To account for proteins targeted to some of the common bacterial hyperstructures and host-destined SCLs, new subcategory localizations have been introduced in PSORTb 3.0, as listed in Table 2.1. This represents, to our knowledge, the first implementation of subcategories for primary SCL localizations, for an SCL predictor. These subcategory localizations for a protein were identified using the SCL-BLAST module, which infers localization by homology using criteria that are of measured high precision (Nair and Rost, 2002). Proteins detected to have a secondary localization are also predicted as one of the four main categories for Gram-positive bacteria or one of five main compartments for Gram-negative bacteria (or similarly for those bacteria with atypical cell structures). Any protein exported past the outer-most layer of the bacterial cell is considered as extracellular, while proteins localized to one of the membranes that are part of a hyperstructure (such as the flagellum) are identified both as an inner or outer membrane protein as well as a protein of that hyperstructure. The basal components of the flagellum are not annotated as such, since they are often homologous to proteins that are not part of the flagellar apparatus (for example, a general ATPase).

Table 2.1 New subcategory subcellular localizations predicted by PSORTb 3.0

| Subcellular localization subcategories | Description |
|---|---|
| Host-Associated | Any proteins destined to the host cell cytoplasm, cell membrane or nucleus by any of the bacterial secretion systems |
| Type III Secretion | Components of the type III secretion apparatus |
| Fimbrial | Components of a bacterial or archaeal fimbrium or pilus |
| Flagellar | Components of a bacterial or archaeal flagellum |
| Spore | Components of a spore |

#### 2.3.2.2 Implementation changes to software

The implementation of the new version of PSORTb is similar to version 2.0 (Gardy *et al.,* 2005), with the following changes: motifs that provided false prediction results were either updated or removed. SCL-BLASTdbs for both Gram-positive and Gram-negative options were updated with the newly expanded dataset. The trans-membrane α-helix predictor module HMMTOP (Tusnady and Simon, 2001) was replaced with S-TMHMM, an open source trans-membrane (TM) α-helix predictor (Viklund and Elofsson, 2004). The program is modified such that the software reports the number of TM-helices predicted. The module is referred to as ModHMM within PSORTb. As with the PSORTb 2.0 set-up, this module first examines if an alpha helix was predicted in the first 70 amino acid residues; if so, this helix would be subtracted. It then examines the rest of the protein sequence, returning a positive prediction for cytoplasmic membrane SCL if more than two TM helices are found, to ensure high precision. Although this leads to membrane-associated proteins being under-predicted by this module, such proteins are instead predicted by the SCL-BLAST module and SVMs (mentioned below).

All SVMs, except for the Gram-negative outer membrane SVM module and Gram-positive cytoplasmic SVM module, were re-trained with the new dataset following the protocols of PSORTb 2.0 paper (Gardy *et al.*, 2005). The aforementioned two SVMs were not updated because the new SVMs did not improve significantly in performance when retrained. For PSORTb 2.0, we made use of an implementation of generalized suffix tree (Wang *et al.*, 1994) to extract frequent subsequences which occur in more than a predefined fraction of total number of proteins of interest. These frequent subsequences were used as features to discriminate localizations of related proteins. The implementation first sampled a subset of related proteins, then extracted frequent subsequences from this subset and finally checked whether these frequent subsequences were frequent in all related proteins. This method may miss some frequent subsequences or produce false positives. To overcome this issue, we used another augmentation of generalized suffix tree (Matias *et al.*, 1998). The algorithm guarantees returning all the frequent subsequences and its running time is in the order of the total length of the related protein sequences.

A Bayesian network combines all module predictions and generates one final localization result based on the performance accuracies of each of the updated modules. Table 2.2 shows the list of modules used in PSORTb 3.0.

Table 2.2 List of modules used and features used for predictions in PSORTb 3.0. Not all modules received an update since version 2.0.

| *Module* | *Features used for prediction* | *SCLs predicted* | *Updated since PSORTb 2.0?* |
|---|---|---|---|
| **Signal peptide prediction** | N-terminal signal peptide | non-cytoplasmic | No |
| **SVMs** | frequent subsequences within protein sequences | all SCLs | Yes |
| **ModHMM** | transmembrane α-helices | CM | Yes |
| **PROSITE motifs** | motifs associated with specific SCLs | all SCLs | Yes |
| **PROFILE motifs** | motifs associated with specific SCLs | all SCLs | No |
| **Outer membrane motifs** | motifs associated with β-barrel OM proteins | OM | No |
| **SCL-BLAST** | homology | all SCLs | Yes |

#### 2.3.2.3 New prediction categories for Archaea and atypical prokaryotic organisms

The SCL predictor for Archaea was implemented with similar components as the Gram-positive predictor, producing predictions for four localizations and two subcategory localizations (flagellum and fimbrium), but using the archaeal training dataset mentioned above. Any motifs that reduced the precision for archaeal SCL prediction were removed.

Two other categories were implemented for bacteria with atypical cellular structures – organisms that stain Gram-positive but have an outer membrane, and organisms that stain Gram-negative but have no outer membrane. For the former category, the Gram-negative pipeline was employed, which enables outer membrane and periplasmic localizations to be predicted. For the latter category, the Gram-positive modules were used, but the cell wall localization prediction was disabled, since the intended organisms (i.e. Tenericutes) lack cell walls.

#### 2.3.2.4 Software usability improvements

To improve usability of the new software version, the web interface of PSORTb 3 now allows user to upload a batch job (such as an entire proteome), and a formatted results file is returned to the user by email when computations are completed. The installation process of the standalone software has also been improved such that the process requires fewer packages and can be installed in a more automated manner. PSORTb 3.0 works with most versions of Linux as well as Mac OS X (except Snow Leopard at press time).

### 2.3.3 Software Evaluations – using literature and Swiss-Prot-based datasets

Five-fold cross validation was performed on the updated Gram-positive bacteria, Gram-negative bacteria and archaeal datasets using the approach as described in the PSORTb 2.0 paper (Gardy *et al*., 2005). In order to use these new datasets to evaluate the performance of other SCL predictors, proteins from the training set of PSORTb 2.0 were subtracted from this evaluation dataset, since this particular set of proteins is included in the training dataset of most of the bacterial SCL prediction tools. To improve the robustness of the assessment of accuracy, homology reduction was performed on the testing datasets using CD-HIT (Li and Godzik, 2006) such that none of the sequences in the testing set exhibited greater than 80% identity with other sequences in the set. Several different threshold values were tested, but we chose 80% as the benchmark threshold in order not to lose functional divergence between the test proteins. Performance metrics used to evaluate different software include precision, defined as TP / (TP + FP); recall, defined as TP / (TP + FN); accuracy, defined as (TP + TN) / (TP + TN + FP + FN); and Matthew's Coefficient Constant (MCC), defined as

$$MCC = \frac{TP \bullet TN - FP \bullet FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{1}$$

The following web servers were benchmarked for their predictive capabilities, in addition to PSORTb versions 2.0 and 3.0: CELLO version 2.0 (Yu *et al.*, 2006) , Gneg-PLoc (Chou and Shen, 2006), Gpos-PLoc (Shen and Chou, 2007), and Proteome Analyst version 2.5 (PA 2.5) (Lu *et al.,* 2004), whose performance was previously shown to be comparable to PSORTb 2.0 (Gardy and Brinkman, 2006). Proteome Analyst 3.0 (PA 3.0), an unpublished method, was also included in this benchmark analysis, though it could only be evaluated using a new proteomics-derived experimental dataset, since we could not confirm that our test data was not in the training data for this software. Methods that are specific to an organism, such as TBPred (Rashid *et al.*, 2007), and methods that do not allow for user submission of protein sequences, such as LocateP (Zhou *et al.*, 2008) and Augur (Billion *et al.*, 2006)), could not be included in this comparison. Two of the recently developed methods, PSL101 (Su *et al.*, 2007) and PSLDoc (Chang *et al.*, 2008) were not tested since the servers could not handle large testing datasets. Once the level of precision was determined for each software, those with highest precision were also evaluated for "proteome coverage", i.e. the proportion of proteins predicted in a deduced proteome from a genome, at that level of precision.

### 2.3.4 Proteomics analysis

We performed a laboratory analysis to construct an experimental dataset of proteins from a Gram-negative bacterium, *Pseudomonas aeruginosa* PA01, which was used to assess PSORTb 2.0, PSORTb 3.0, PA 2.5, and PA 3.0. This represents an independent dataset that includes hypothetical and uncharacterized proteins with previously unknown subcellular localizations. *P. aeruginosa* is a bacterium noted for its diverse metabolic capacity and large genome/proteome size, and so represents an excellent organism with which to generate such a dataset (Stover *et al.,* 2000). To generate this experimental dataset, we extracted protein samples from the cytoplasmic, periplasmic and secreted fractions of *P. aeruginosa* PA01. The resulting proteins in each fraction were digested to peptides and differentially labeled using formaldehyde isotopologues (Chan and Foster, 2008) prior to analysis by LC-MS/MS, exactly as previously described (Chan *et al.*, 2006). Abundance ratios between SCL were calculated using MSQuant (http://msquant.sourceforge.net/). To ensure a high-quality dataset with minimal contaminating proteins from other subcellular compartments, proteins that were only found in the cytoplasmic fraction and never in the other two soluble fractions were used to assess PSORTb 3.0 and PA 3.0 prediction results. This dataset was also felt to be most appropriate for assessment, since our analysis had suggested that most proteins of previously unknown localization in the old version of PSORTb were most likely cytoplasmic proteins. Further details on the experimental protocols for this proteomics analysis of the subcellular fractions are described in the next section.

### 2.3.5 Mass Spectometry protein identification

#### 2.3.5.1 Sample preparation

Protein concentrations of samples from secreted and cytoplasmic fractions of *Pseudomonas aeruginosa PAO1* extracted from mid-log growth phase were measured using Coomassie Plus Protein Assay reagent (Pierce). 50 µg of proteins were then taken from each fraction, precipitated using ethanol/sodium acetate (Foster *et al.,* 2007), digested with trypsin in solution (Silverman *et al.,* 2008) and labeled by formaldehyde reductive dimethylation as described in (Hsu *et al.*, 2003). The secreted fraction was labeled with deuterated formaldehyde and mixed at a 4 to 1 ratio with the cytoplasmic fraction, which was labeled with regular formaldehyde. Labeled and digested samples were then subjected to subfractionation, using either MicroRotofor (BioRad) or Off-Gel (Agilent) isoelectric focusing (IEF) methods (Sliverman *et al,* 2008). STop and Go Extraction (STAGE) tips were used in between most steps to desalt, purify and concentrate the peptides (Rappsilber *et al,* 2003). The fractionated peptides were dried in vacuum concentrator (Eppendorf), resuspended in sample buffer (1% trifluoroacetic acid, 0.5% acetic acid, and 3% acetonitrile), and analyzed using either LTQ-Orbitrap (Thermo Electron) or LTQ-FT (Thermo Electron, Bremen, Germany) with settings as described in (Chan *et al.*, 2006). The periplasmic fraction was processed, mixed with differentially labeled cytoplasmic fraction and analyzed with the mass spectrometer in a similar fashion as the secreted fraction.

#### 2.3.5.2 Protein identification

Protein identification consisted of searching measured MS/MS fragment spectra against a database of theoretical mass spectra predicted from a protein sequence database.

DTASuperCharge (http://msquant.sourceforge.net/) was used to correct monoisotopic peak assignments. The commercial software Mascot (Matrix Science) was used to search against the *in silico* trypsin-digested *Pseudomonas aeruginosa* PAO1 proteome as well as porcine trypsin and human keratin sequences. The search criteria used were as follows: trypsin cleavage specificity with up to one missed cleavage, cysteine carbamidomethyl fixed modification, deuterated and regular dimethylated lysine and N-terminus variable modifications, ±5ppm peptide tolerance, ±0.6Da MS/MS tolerance, and ESI-Trap fragmentation scoring. The reversed version of the *in silico* trypsin-digested *Pseudomonas aeruginosa* PAO1 proteome database was used at the final step to estimate the false positive rate. A custom-written Perl script pickletrimmer.pl (http://www.chibi.ubc.ca/faculty/foster/softwares) was used to remove mass spectra the peak list files (in Mascot Generic Format) that did not match to any peptides in the database. The trimmed peak list files were combined and subjected to a final Mascot search. Proteins identified by >1 unique peptide in three replicates that are above a score of 15 were considered for quantitation.

#### 2.3.5.3 Quantitation Step

The abundance of a peptide is proportional to the extracted ion chromatogram (XIC), the area under a peak generated by a time-course sampling mass spectra signal intensities as a peptide elutes from the LC column. We used MSQuant (http://msquant.sourceforge.net) to parse Mascot results, recalibrate mass measurements, and to extract quantitative ratios.

## 2.4 Results

### 2.4.1 PSORTb 3.0: Expanded predictive capabilities for all prokaryotes and localization subcategories

We present version 3.0 of PSORTb. Like the version 2 series, version 3.0 has the capability to make predictions for all Bacteria, but now makes predictions for Archaea and bacteria with atypical cell wall/membrane structures as well. Users must simply select the Domain of life (Bacteria or Archaea) and, in the case of bacteria, select whether the organism is Gram-positive or Gram-negative or "Advanced" (i.e. Gram-positive with an outer membrane or Gram-negative without an outer membrane). Localization predictions now include a sub-categorization (see Methods as well as Table 2.1) for more precise identification of localizations (i.e. a protein may be in the outer membrane but also be a component of the flagellar machinery, so it would be classified as "outer membrane", with a subcategory classification as "flagellar").

### 2.4.2 PSORTb 3.0 outperforms PSORTb 2.0 and other SCL prediction tools in terms of precision and recall for bacterial proteins

The overall performance for PSORTb 3.0, calculated using five-fold cross validation, along with the performance of other recently published bacterial SCL prediction tools tested using the homology-reduced dataset, are shown in Table 2.3. The SCL-specific performance values for each predictor can be found in Appendix A - *Supplementary Data* Tables 1 and 2. For the Gram-positive option, both PSORTb 3.0 and PSORTb 2.0 exhibit precision values above 97%, while CELLO 2.5, Gpos-PLoc, and PA 2.5 measured below 95%. Overall recall values were above 90% for all benchmarked software except for PA 2.5, which seems to have an especially low recall (11.5%) for membrane proteins. For the Gram-negative option, PSORTb 3.0 still maintains the highest precision of 97.3% and the highest recall of 94.1%, where recall has increased by 8.8% compared to PSORTb 2.0. PA 2.5, which was previously shown to be comparable to PSORTb 2.0, still exhibits comparable precision (97.3%) and recall (92.0%) with this new test dataset. Although SubcellPredict and SLP-Local also show high overall precision and recall values, their precision values for the periplasmic localization prediction are under 55%. Gneg-PLoc and CELLO 2.5, having precision values below 90%, also exhibit lower specificities for periplasmic localizations (56.5% and 35.2% respectively) as well as outer membrane localizations (66.4% and 34.6% respectively). Overall, PSORTb 3.0 appears to be the most accurate versus all other comparable methods that were tested. Compared to PSORTb 2.0, PSORTb 3.0 appears to predict more cytoplasmic proteins in particular, reflecting difficulty in identifying the localization of such proteins without an improved training dataset (since they have no signals to transport them to other localizations that may be detected). PSORTb 3.0 has a marked improvement over PSORTb 2.0 in recall in particular for Gram-negatives, representing a significant improvement in predictive capability for the only SCL predictor of its kind that is freely available as a standalone package.

Table 2.3 Performance comparisons for Gram-positive and Gram-negative bacterial SCL prediction software

| Software§ | Precision† | Recall† | Accuracy† | MCC† |
|---|---|---|---|---|
| **Gram-positive** | | | | |
| PSORTb 3.0 | 98.2 | 93.1 | 97.9 | 0.79 |
| PSORTb 2.0 | 97.0 | 90.0 | 96.8 | 0.76 |
| CELLO 2.5 | 93.7 | 93.7 | 96.9 | 0.76 |
| Gpos-PLoc†† | 91.2 | 90.7 | 95.5 | 0.64 |
| PA 2.5††† | 90.0 | 81.8 | 90.9 | 0.57 |
| **Gram-negative** | | | | |
| PSORTb 3.0 | 97.3 | 94.1 | 98.3 | 0.85 |
| PA 2.5 | 97.3 | 92.0 | 97.9 | 0.85 |
| PSORTb 2.0 | 95.9 | 85.3 | 96.3 | 0.69 |
| SubcellPredict* | 94.3 | 94.3 | 96.0 | 0.52 |
| SLP-Local* | 93.8 | 93.8 | 95.9 | 0.59 |
| Gneg-PLoc** | 89.6 | 88.9 | 95.7 | 0.65 |
| CELLO 2.5 | 87.5 | 87.5 | 95.0 | 0.61 |

§PA 3.0 is not included in the analysis since we are unable to determine the degree of overlap between our testing dataset and the training dataset of PA 3.0.

† Precision = TP / (TP + FP); Recall = TP / (TP + FN); Accuracy = (TP + TN) / (TP + FP + TN + FN);

$$MCC = \frac{TP \bullet TN - FP \bullet FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP = # true positives, FP = # false positives, TN = # true negatives, FN = # false negatives, MCC = Matthew's Coefficient Constant

†† Software also predicts periplasmic SCL. None of the testing dataset proteins received a periplasmic SCL prediction.

††† Software only predicts cytoplasmic, membrane, and extracellular categories. All proteins (including cell wall proteins) submitted to the server will receive one or more of these 3 localization predictions (or 'No predictions').

*Software only predicts cytoplasmic, periplasmic, and extracellular categories. All proteins (including membrane proteins) submitted to the server will receive one of these 3 SCL predictions

** Software also predicts flagellar, fimbrium and nucleoid localizations; however, none of test dataset proteins received one of these 3 SCL predictions

### 2.4.3 PSORTb 3.0 outperforms PSORTb 2.0 and other bacterial prediction software for predicting archaeal SCLs

The domain of Archaea exhibits highly diverse morphologies. However, for most archaeal organisms, the basic compartments are similar to that of monodermic bacteria, namely: cytoplasmic space, cell membrane, a proteinaceous cell wall, and secreted proteins. The five-fold cross validation results for the archaeal predictor are shown in Table 2.4. SCL-specific performance values for different predictors can be found in Appendix A - *Supplementary Data* Table 3. We compared the performance of our archaeal-specific predictor to Gram-positive bacterial SCL predictors since there is no other archaeal-specific predictor. We found that overall, Gram-positive bacterial predictors can predict archaeal cytoplasmic and membrane proteins with relatively high recall and precision, but with extracellular proteins the precision is quite low. PSORTb 3.0 is able to capture predictions for some of the archaeal-specific proteins and demonstrates superiority in performance compared to PSORTb 2.0 and to other Gram-positive bacterial SCL predictors, and now represents the first predictor specifically designed for the important domain of Archaea.

Table 2.4 Performance comparison for archaeal proteins between PSORTb 3.0 archaeal option and software with Gram-positive SCL prediction capability

| Software§ | Precision | Recall | Accuracy | MCC† |
|---|---|---|---|---|
| PSORTb 3.0 | 97.2 | 93.4 | 97.7 | 0.83 |
| PSORTb 2.0 | 95.7 | 81.0 | 94.3 | 0.59 |
| Gpos-PLoc* | 92.3 | 92.3 | 96.2 | 0.65 |
| PA 2.5** | 90.0 | 77.5 | 89.6 | 0.38 |
| CELLO 2.5 | 86.5 | 86.5 | 93.2 | 0.46 |

§PA 3.0 is not included in the analysis since the exact content of the training dataset is unknown and may skew the cross-validation results.

†See Table 2.3 footnotes for definitions of the four performance metrics.

*Software also predicts periplasmic SCL. None of the testing dataset proteins received a periplasmic SCL prediction.

**Software does not predict cell wall localization.

### 2.4.4 Evaluation of PSORTb and PA 3.0 using a new proteomics-derived experimental dataset – PSORTb 3.0 has highest recall

PA 3.0, an unpublished version of bacterial SCL predictor is also available through the Proteome Analyst website (http://webdocs.cs.ualberta.ca/~bioinfo/PA/) with updated algorithms. We wished to compare the accuracy of this predictor, but we were unable to determine the content of the software's training dataset and the degree of overlap with our testing dataset. To account for the bias associated with testing and training with the same dataset, we therefore opted to evaluate PSORTb 3.0 and PA 3.0 using an independent dataset of 171 cytoplasmic proteins from the Gram-negative organism *P. aeruginosa* PA01. This dataset likely contains some proteins that are part of the training dataset of one or both tools, but most of the proteins with unknown functions that are identified from the experiment were never previously characterized for their localizations before and would not have been included in any SCL predictor's training data. This experimentally-generated proteomics dataset should more accurately evaluate the software's predictive capabilities for analyzing a proteome. Table 2.5 shows the precision and recall of each predictor, where a false positive is defined as a protein receiving an SCL prediction that is not "cytoplasmic". The prediction results for PSORTb 2.0 and PA 2.5 are also shown for reference. Similar to the results derived using the literature-derived dataset, PSORTb 3.0 and PA 3.0 demonstrate higher precision and recall compared to PSORTb 2.0 and PA 2.5. However, more proteins in this dataset receive a prediction from PSORTb 3.0 than from PA 3.0, indicating that PSORTb 3.0 achieves higher recall than PA 3.0. A full list of proteins used in the proteomic analysis, and their prediction results, can be found in Appendix A - *Supplementary data* Table 4.

Table 2.5 Evaluation of PSORTb 3.0, PSORTb 2, PA 3.0 and PA 2.5 using LC-MS proteomics dataset of proteins found exclusively in the cytoplasmic fraction when comparing to the periplasmic and extracellular fractions of *Pseudomonas aeruginosa* PA01

| Software | Precision* | Recall |
|---|---|---|
| PSORTb 3.0 | 96.3 | 91.8 |
| PA 3.0 | 95.9 | 81.3 |
| PA 2.5 | 90.7 | 51.5 |
| PSORTb 2.5 | 90.3 | 54.4 |

* Precision in this case refers to TP/(TP + FP), where FP refers to proteins predicted as SCLs other than "Cytoplasmic" or "Unknown".

### 2.4.5 Proteome prediction coverage is increased

Although PSORTb 3.0 exhibited higher recall compared to PSORTb 2, our main goal was to increase prediction coverage for whole bacterial proteomes while maintaining a high level of precision. Figure 2.1 shows the coverage results of PSORTb 2.0 compared to PSORTb 3.0. Coverage is defined as the proportion of proteins in a deduced proteome that receives a prediction from the software at a measured level of precision (see above for precision results). The proteomes analyzed were chosen to cover a wide spectrum of bacterial phyla, ranging from well-studied model organisms such as *Escherichia coli* to lesser studied species that previously had low predictive coverage with PSORTb 2.0. Among the species tested, on average there was a 17.1% increase in proteome prediction coverage for Gram-negative bacterial proteomes and 5.9% increase for Gram-positive bacterial proteomes. Among the selected Gram-negatives, the *Aquifex aeolicus* proteome achieved the highest coverage (90.5%). *Helicobacter pylori* obtained the highest coverage increase (23.9%) while *P. aeruginosa* PA01 only gained 10.6% of coverage increase, the lowest of the Gram-negative list. *Lactobacillus johnsonii*, among the list of tested Gram-positive organisms, gained 7.9% in coverage, while *Clostridium difficile* received a modest boost of 2.8% in predictive coverage. Overall, proteome prediction for all tested organisms benefitted from the performance boost from PSORTb 3.

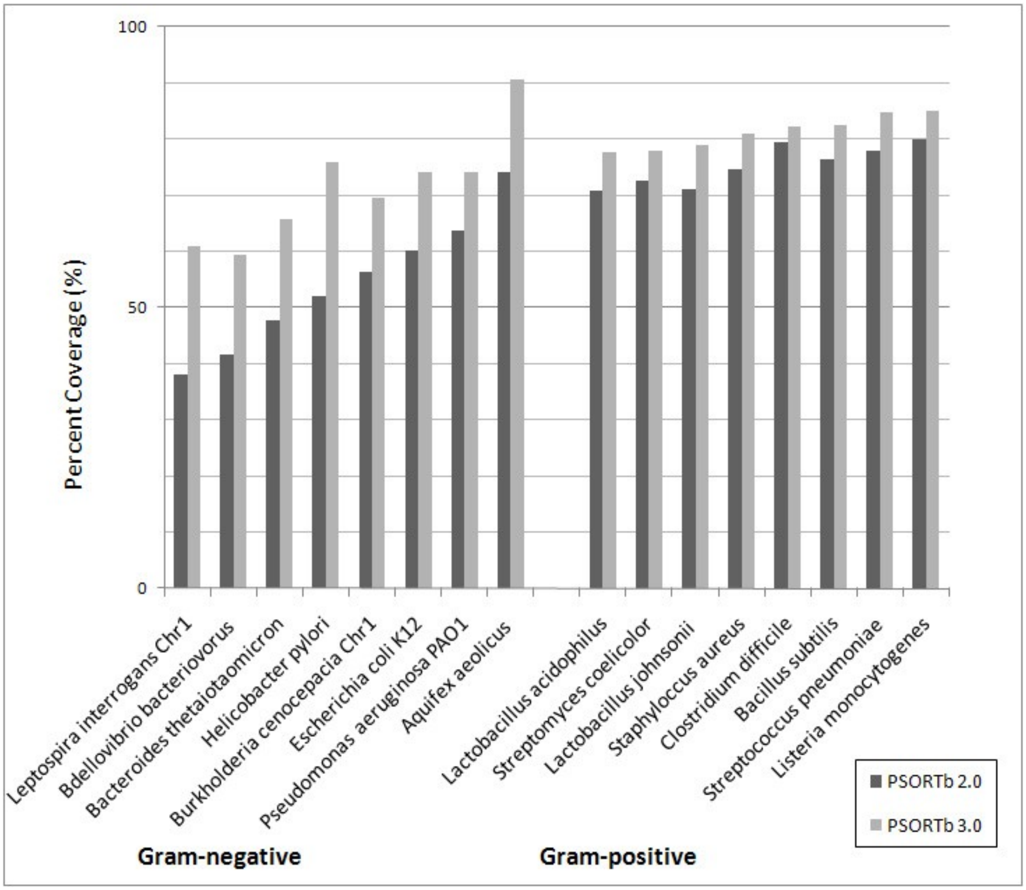


Figure 2.1 Genome coverage prediction for PSORTb 2.0 and PSORTb 3.0 for Gram-negative and Gram-positive bacteria genomes. Chr1 denotes chromosome 1.

### 2.4.6 PSORTb 3.0 and PA 3.0 make complementary predictions – a combined analysis with both methods has the highest coverage overall

Since PA 3.0 was the only comparable program to PSORTb in terms of precision and ability to not force predictions (i.e. have an "unknown" prediction category), and has been validated using the proteomics dataset to have better performance compared to PA 2.5, we examined the prediction results for combining PSORTb 3.0 and PA 3.0. We tested this on several Gram-positive and Gram-negative bacterial genomes, including both model organisms and lesser-studied species. The results are shown in Figure 2.2. In combination, the two predictors were capable of generating predictions for about 80-95% of all bacterial proteins encoded in the selected bacterial genomes, which exhibits an impressive increase versus the previous predictive capability of PSORTb 2.0 (57-75%) (Gardy *et al.,* 2005) and PA 2.5 (67-76%) (Lu *et al.,* 2004). On average, 52.5% of the proteins in each genome-derived proteome received consensus SCL predictions from the two predictors. About 20-30% of the genes were predicted by either PA 3.0 or PSORTb 3.0 but not both programs, which shows a significant level of complementarities for two very precise predictors. Of the cases with different predictions (5-10%), we found that over half of these predictions consist of neighbouring localizations (e.g. cytoplasmic vs. cytoplasmic membrane). Upon manual inspection, these likely reflect the nature of peripheral membrane proteins that could not be detected as such by each predictor alone. For example, one program predicted cytoplasmic and the other predicted cytoplasmic membrane. For such membrane-associated proteins, technically both programs could be considered correct. For Gram-positive bacterial proteomes, although PA 3.0 does not predict a "cell wall" SCL, many of the PSORTb-predicted cell wall proteins received “membrane” or “extracellular” predictions by PA, which does reflect the fact that many of them are membrane anchored and protrude into the extracellular space. Taking these points into consideration, only roughly 2.5-5% of the predictions appear to disagree, reflecting an expected level of error given the precision of each method. Taken together, it appears that combining the two methods notably increases genome prediction coverage indicating that the two methods are complementary and should be used together when possible.

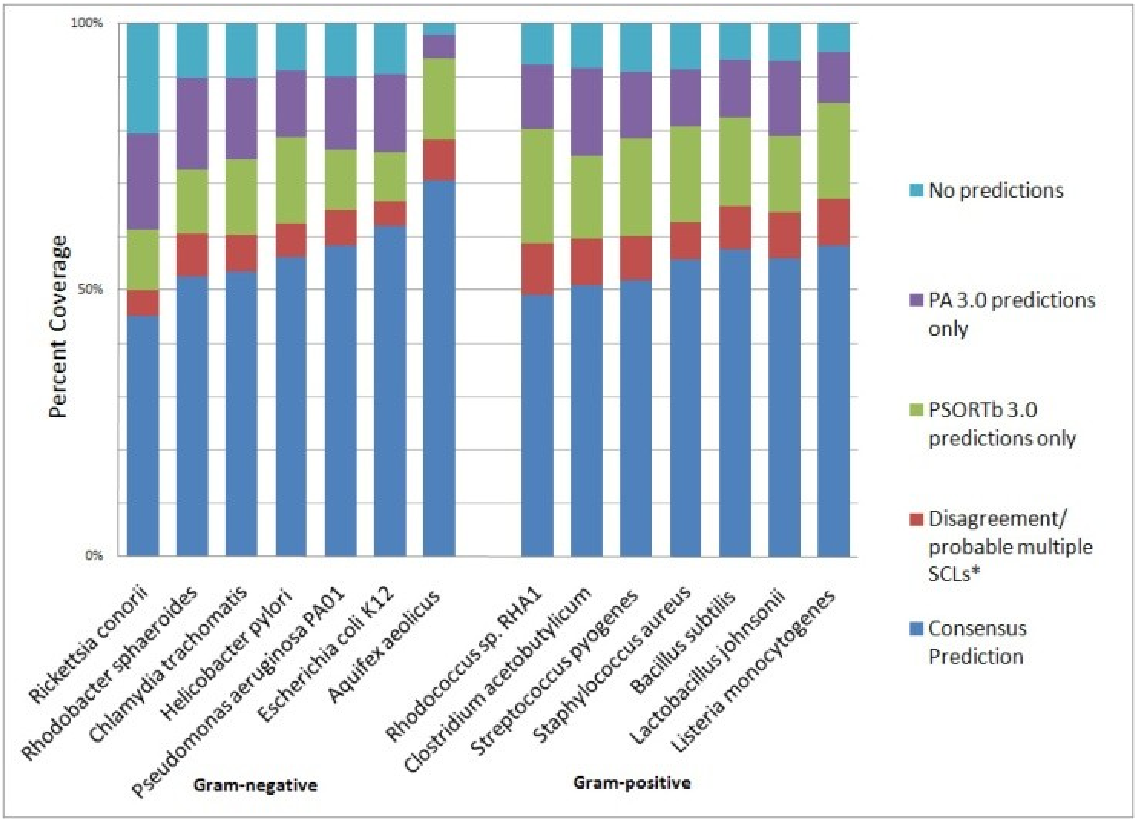


Figure 2.2 Genome prediction coverage results from combining PSORTb 3.0 and PA 3.0 output. The majority of the “disagreement” cases are boundary localizations (membrane prediction and a neighboring compartment). This likely reflects the true nature of the proteins. Only a small fraction of the disagreements (2.5-5% of the deduced proteome) are non-boundary cases.

## 2.5 Discussion

The new version of PSORTb was created with the following improvements in mind: refining localization prediction, implementing archaeal SCL prediction capabilities, increasing software recall and proteome prediction coverage while maintaining high precision, and ensuring user-friendly software installation as well as usage. We found it necessary to implement subcategory localizations for several reasons. First of all, we have anecdotally observed that effector proteins secreted by the type III and type IV secretion systems were predicted as cytoplasmic proteins by the PSORTb 2 and most other SCL predictors due to the fact that their final destination is the host cell cytoplasm and likely contain properties similar to cytoplasmic proteins. Secondly, for structural proteins that are parts of a bacterial organellar apparatus, it would be more informative to note the apparatus itself as localization in addition to the main subcellular compartment currently assigned by PSORTb. Although the initial BLAST-based approach may be limited in capturing only effector proteins with enough sequence similarity to each other, we hope to further expand the dataset of effector proteins for training as they are identified. Having a subcategory localization detection allows PSORTb to give these types of proteins a more refined localization annotation, for example: "extracellular – T3SS (type III secretion apparatus)" rather than just the misleading classification of "extracellular".

We have built the first SCL predictor specific for the domain of Archaea and assessed its performance with a dataset of archaeal proteins. Although Gram-positive bacterial predictors seem to perform quite well for archaeal cytoplasmic and membrane proteins, the low recall values show that a bacterial-only training dataset fails to predict archaeal cell wall and extracellular proteins well. Because of the unique nature of archaeal cell walls, which usually consist of a proteinaceous S-layer rather than peptidoglycan found in bacteria, proteins that reside in this localization can be quite different from cell wall proteins of Gram-positive organisms. If the training dataset does not contain representative properties for its localization category, no software would be able to generate highly accurate predictions for that particular category or that particular species. To further improve predictions for archaeal proteomes, we suspect that a more extensive training dataset needs to be collected for cell wall and secreted proteins in particular.

We have also added the capability to handle predictions for the four possible different types of bacteria – Gram-positive with and without an outer membrane, and Gram-negative with and without an outer membrane. As the diversity of bacteria being studied increases through metagenomics and other larger scale studies, having such capability to handle the diversity found in this domain of life will become increasingly important. Future research should focus on increasing the ability of SCL predictors to handle more specialized types of bacteria and archaea with atypical cell structures.

Most high-throughput mass spectrometry-based proteomic studies of subcellular fractions tend to include proteins from other subcellular compartments, due to some degree of cell lysis (Rey *et al.*, 2005). We were able to generate a relatively small dataset of highly-reliant cytoplasmic identifications by eliminating any proteins that were found also in periplasmic or extracellular fractions in a proteome-scale analysis. While this approach will miss a lot of potential cytoplasmic proteins, this dataset is of high specificity and contains proteins that are not part of any SCL predictor's training dataset. For the other localizations, however, it is much more difficult to obtain relatively contaminant-free fraction samples, due to the fact that highly abundant cytoplasmic proteins (such as ribosomal proteins and molecular chaperone GroEL) tend to contaminate other fractions at such high levels. Further improvements in protein sample preparation for the non-cytoplasmic fractions are needed if we want to use this approach to validate software precision for the other SCLs.

We show that with the addition of new training data, PSORTb's recall and coverage improved and the performance remains ahead of other comparable bacterial SCL prediction software. This demonstrates that the effect of increasing training data size on improving such a prediction tool is still an effective way to increase predictive accuracy. By combining PA and PSORTb, two of the most accurate SCL predictors, we can now predict localizations for 80-95% of most bacterial proteomes. Efforts to further improve prediction capabilities should focus on developing approaches to tackle the last 5-20% of the proteomes. Preliminary analysis suggests that these are likely to be uncharacterized genes that are either common to a smaller subset of prokaryotic classes or unique to particular strains. A combined effort of small-scale as well as refined high-throughput experimental approaches, continual data mining from literature, and algorithm improvement will be required to determine the localization of these proteins. The significant number of cases where PSORTb and PA predicted localizations to neighboring compartments highlights the need to further refine the SCL classification and identification of peripheral membrane proteins, which include proteins attached to the inner or outer membrane via a single alpha helix, a lipid moiety, or covalently linked to an integral membrane protein. Although LocateP and Augur begin to deal with this issue, such refinement should eventually be incorporated into whole-genome SCL analyzing software. For example, currently lipoproteins are detected by the SVMs and the SCL-BLAST modules, but with better understanding of lipoprotein motifs and general properties, future SCL predictors may generate more precise predictions with a "lipoprotein" subcategory identification for this class of proteins.

## 2.6 Conclusion

In summary, PSORTb 3.0 continues to be the most precise SCL predictor of its kind and now has notably increased recall and predictive coverage. It is also the most flexible SCL prediction software for prokaryotes, with both an online web server (with associated email client for larger jobs) as well as an open source standalone version with simplified installation procedure, which allows it to be easily used locally or incorporated into any existing bioinformatics analysis pipeline. With the added predictive capability of archaeal protein SCL prediction, predictions for bacteria with atypical cell morphologies, and the addition of new predictive subcategories, this represents the first SCL predictor designed to handle a diverse range of all prokaryotes and handle prokaryotic subcategory localizations. Our results show that this tool can be effectively complemented by PA 3.0, generating an impressively high number of SCL predictions for proteomes at high precision. This new version of PSORTb, as well as the datasets used to train the software, will serve as a useful resource for bioinformaticists and the greater microbiology community.

## 2.7 Acknowledgements

The authors would like to thank Francis Lim for protein sample processing for mass spectrometry analysis, as well as Kurt McMillan and Yifeng Liu for providing PA 3.0 whole genome prediction results as well as helpful discussions. We thank Shannan Ho Sui for constructive feedback. NYY was supported in part by a training grant from the BC Proteomics Network. JRW was supported by a Postgraduate Scholarship from the Natural Sciences and Engineering Research Council of Canada, and by the Canadian Institutes of Health Research and Michael Smith Foundation for Health Research Strategic Training Program in Bioinformatics Graduate Award. LJF is the Canada Research Chair in Quantitative Proteomics. LJF and FSLB are Michael Smith Foundation for Health Research Scholar and Senior Scholar, respectively. FSLB is also a Canadian Institutes of Health Research New Investigator.

*Funding*: This work was funded primarily by the Natural Sciences and Engineering Research Council of Canada, with the support of the Cystic Fibrosis Foundation and Simon Fraser University Community Trust.

## 2.8 References

Arnold, R. et al. (2009) Sequence-based prediction of type III secreted proteins. PLoS Pathog. 5, e1000376.

Billion, A. et al. (2006) Augur--a computational pipeline for whole genome microbial surface protein prediction and classification. Bioinformatics, 22, 2819-2820.

Bulashevska, A. and Eils, R. (2006) Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. BMC Bioinformatics, 7, 298.

Chan, Q.W. et al. (2006) Quantitative comparison of caste differences in honeybee hemolymph. Mol. Cell. Proteomics., 5, 2252-2262.

Chan, Q.W. and Foster, L.J. (2008) Changes in protein expression during honey bee larval development. Genome Biol., 9, R156.

Chang, J.M. et al. (2008) PSLDoc: Protein subcellular localization prediction based on gapped-dipeptides and probabilistic latent semantic analysis. Proteins, 72, 693-710.

Chou, K.C. and Shen, H.B. (2006) Large-scale predictions of gram-negative bacterial protein subcellular locations. J. Proteome Res., 5, 3420-3428.

de Champdoré, M. et al. (2007) Proteins from extremophiles as stable tools for advanced biotechnological applications of high social interest. J. R. Soc. Interface, 4,183-191.

Foster, L.J. *et al*. (2003) Unbiased quantitative proteomics of lipid rafts reveals high specificity for signaling factors. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 5813 –5818.

Gardy, J.L. et al. (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics, 21, 617-623.

Gardy, J.L. and Brinkman, F.S.L. (2006) Methods for predicting bacterial protein subcellular localization. Nat Rev Micro, 4, 741-751.

Hsu, J.L. *et al.* (2003). Stable-isotope dimethyl labeling for quantitative proteomics. *Anal Chem* **75**,6843-52.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22, 1658-1659.

Matias. et al. (1998) Augmenting Suffix Trees with Applications. In ESA 1998, pp. 67-78.

Matsuda, S. et al. (2005) A novel representation of protein sequences for prediction of subcellular location using support vector machines. Protein Sci., 14, 2804-2813.

Melli, G. et al. (2007) Recognition of Multi-sentence n-ary Subcellular Localization Mentions in Biomedical Abstracts. In Proceedings of LBM-2007.

Miyata, M and Ogaki, H. (2006) Cytoskeleton of mollicutes. J. Mol. Microbiol. Biotechnol., 11, 256-264.

Nair, R. and Rost, B. (2002) Sequence conserved for subcellular localization. Protein Sci., 11, 2836-2847.

Niu, B. et al. (2008) Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins. Mol.Divers., 12, 41-45.

Rappsilber, J. *et al.* (2003) STop And Go Extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75**,663-670.

Rashid, M. et al. (2007) Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. BMC Bioinformatics, 8, 337.

Rey, S. et al. (2005) Assessing the precision of high-throughput computational and laboratory approaches for the genome-wide identification of protein subcellular localization in bacteria. BMC Genomics, 6, 162.

Rey, S. et al. (2005) PSORTdb: a protein subcellular localization database for bacteria. Nucleic Acids Res., 33, D164-168.

Samudrala, R. et al. (2009) Accurate prediction of secreted substrates and identifica-tion of a conserved putative secretion signal for type III secretion systems. PLoS Pathog., 5, e1000375.

Shen, H.B. and Chou, K.C. (2007) Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng.Des.Sel., 20, 39-46.

Silverman, J.M. *et al.* (2008) Proteomics analysis of the secretome of *Leishmania donovani*. *Genome Biology* **9**,R35.

Stover, C.K. et al. (2000) Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. Nature, 406, 959-964.

Su, E.C. et al. (2007) Protein subcellular localization prediction based on compart-ment-specific features and structure conservation. BMC Bioinformatics, 8, 330.

Thompson, B.G. and Murray, R.G. (1981) Isolation and characterization of the plasma membrane and the outer membrane of Deinococcus radiodurans strain Sark. Can. J. Microbiol.,27, 729-734.

Tusnady, G.E. and Simon, I. (2001) The HMMTOP transmembrane topology predic-tion server. Bioinformatics, 17, 849-850.

Viklund, H. and Elofsson, A. (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. Protein Sci., 13, 1908-1917.

Wang, J. et al. (1994) Combinatorial Pattern Discovery for Scientific Data: Some Preliminary Results. In SIGMOD 1994, pp. 115-125.

Winsor, G.L. et al. (2008) Pseudomonas Genome Database: facilitating user-friendly, comprehensive comparisons of microbial genomes. Nucleic Acids Res., 37, D483-488.

Wu, C.H. et al. (2006) The Universal Protein Resource (Uniprot): an expanding universe of protein information. Nucleic Acids Res., 34, D187-191.

Wu, D. et al. (2009) A phylogeny-driven genomic encyclopedia of Bacteria and Archaea. Nature, 462, 1056-1060.

Yu, C.S. et al. (2006) Prediction of protein subcellular localization. Proteins, 64, 643-651.

Zhou, M., et al. (2008) LocateP: genome-scale subcellular-location predictor for bacterial proteins. BMC Bioinformatics, 9, 173.

# 3: PSORTdb – an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea

**Published in**: Nucleic Acids Research 2011, 30(Database issue): D241-D244.

**Author list:** Nancy Y. Yu, Matthew R. Laird, Cory Spencer, Fiona S. L. Brinkman

**Author contributions:** NYY and FSLB co-designed the new database, its web interface and novel interaction capabilities, co-developed the outer membrane (OM) identification procedure, tested and debugged the database, and wrote the manuscript. CS ran the OM identification analysis and developed the web interface for the database. MRL and CS co-developed the database.

## 3.1 Abstract

The subcellular localization (SCL) of a microbial protein provides clues about its function, its suitability as a drug, vaccine or diagnostic target and aids experimental design. The first version of PSORTdb provided a valuable resource comprising a data set of proteins of known SCL (ePSORTdb) as well as pre-computed SCL predictions for proteomes derived from complete bacterial genomes (cPSORTdb). PSORTdb 2.0 (http://db.psort.org) extends user-friendly functionalities, significantly expands ePSORTdb and now contains pre-computed SCL predictions for all prokaryotes—including Archaea and Bacteria with atypical cell wall/membrane structures. cPSORTdb uses the latest version of the SCL predictor PSORTb (version 3.0), with higher genome prediction coverage and functional improvements over PSORTb 2.0, which has been the most precise bacterial SCL predictor available. PSORTdb 2.0 is the first microbial protein SCL database reported to have an automatic updating mechanism to regularly generate SCL predictions for deduced proteomes of newly sequenced prokaryotic organisms. This updating approach uses a novel sequence analysis we developed that detects whether the microbe being analyzed has an outer membrane. This identification of membrane structure permits appropriate SCL prediction in an auto-updated fashion and allows PSORTdb to serve as a practical resource for genome annotation and prokaryotic research.

## 3.2 Introduction

Protein subcellular localization (SCL) prediction aids inference of protein function, identifies candidates for drug or vaccine targets, reveals suitable targets for microbial diagnostics and provides directions for experimental design. For biomedical applications, identification of cell surface and secreted proteins from pathogenic bacteria may lead to the discovery of novel therapeutic targets. Characterizing cell surface and extracellular proteins associated with non-pathogenic Bacteria and Archaea can have industrial uses, or play a role in environmental detection.

The first SCL prediction software, called PSORT, was developed in 1991 by Kenta Nakai for bacteria, animals and plants (1, 2). PSORT II, iPSORT and WoLF PSORT were subsequently developed for eukaryotic species (3–5). PSORTb and PSORTb 2.0 were later developed in 2003 and 2005 specifically for Gram-negative and -positive bacterial protein SCL prediction, with a focus on high-precision/specificity predictions (6,7). They have been the most precise SCL prediction methods developed (8). However, recently PSORTb version 3.0 was developed, with 98% precision for Gram-positive bacteria and 97% precision for Gram-negative bacteria, surpassing PSORTb 2.0 and other available prokaryotic SCL predictors (9). PSORTb 3.0 also provides improved genome prediction coverage (higher recall at high precision), as well as the ability to predict a broader range of prokaryotes including Archaea as well as bacteria with atypical membrane/cell wall structures. In addition, PSORTb 3.0 now identifies subcategory localizations for proteins destined to specialized bacterial organelles (such as the flagellum and pilus) as well as host cell destinations.

The speed in which prokaryotic genomes are sequenced has been increasing at a dramatic rate thanks to the availability of sequencing technologies which can decode DNA sequences at a dramatically increased throughput with lower cost. This creates a challenge for maintaining up-to-date functional annotation of these newly sequenced genomes (10). Given the high accuracy of computational SCL prediction for prokaryotes (8), some genome annotation groups have incorporated SCL prediction into their bioinformatics annotation pipeline (11). Instead of having many different researchers compute the same prokaryotic protein SCL prediction repeatedly when needed, it would be more efficient to create a centralized database of pre-computed SCL prediction results that is continually updated to incorporate SCL predictions for newly sequenced organisms.

Several databases containing prokaryotic SCL information have been developed over the years (see http://www.psort.org for a list), such as DBSubLoc, PA-GOSUB and UniProt (12–14). Some are developed specifically for certain types of bacteria; for example, LocateP Database and Augur contain localization predictions specific to Gram-positive bacteria (15,16); others like DBMLoc are specific for multiple SCLs (17). Some incorporate predictions from multiple SCL-prediction tools like CoBaltDB (18). However, none of them are reported, or observed, to be continually updated in a frequent, regular fashion to accommodate newly sequenced genomes, nor do they contain high-precision predictions suitable for handling diverse prokaryotic cellular structures.

PSORTdb (19) is a database initially developed in 2005 to contain experimentally determined (ePSORTdb) and computationally predicted (cPSORTdb) protein SCLs for Bacteria. The computational predictions in cPSORTdb were originally generated by

PSORTb 2.0, the most precise bacterial SCL predictor of its time (7). It is widely used by researchers wishing to identify the SCL of specific proteins, verify high-throughput experimental results, as well as those who need a training data set to develop novel SCL prediction software.

To keep up with the increasing rate of prokaryotic genomes sequenced, and a new version of PSORTb (version 3.0), we have now developed a new version of PSORTdb (version 2.0; http://db.psort.org) that automatically computes PSORTb 3.0 SCL prediction results as new prokaryotic genomes become available through NCBI each month. A largely expanded training data set of proteins with known SCLs have also been added to ePSORTdb. We have also improved the user interface to facilitate easier browsing and searching of the database, as well as convenient download options for filtered search results. Old PSORTb 2.0 prediction results are still maintained in this database for archival purposes. The entire database is freely available through the web or for download, and highlighted features are briefly described below.

## 3.3 New database with largely expanded content

ePSORTdb has significantly expanded and now contains over 9100 entries for Gram-negative bacteria, 2980 entries for Gram-positive bacteria and 800 entries for archaeal proteins (previously ePSORTdb contained a total of only 2171 bacterial proteins). This data set came from manual literature search as well as Swiss-Prot annotations as described in the paper describing PSORTb 3.0 (9). The data set can be used by SCL-prediction software developers to test novel machine algorithms and build better SCL prediction tools. For computationally predicted SCLs in cPSORTdb, to date more than 1000 proteomes deduced from sequenced prokaryotic genomes have been analyzed and the results are available for access on the web server as well as for download. At the time of writing, cPSORTdb contains 1286 Gram-negative bacterial replicons, 508 Gram-positive bacterial replicons, 126 archaeal replicons, 30 replicons belonging to Gram-negative bacteria without outer membrane and 11 replicons that belong to Gram-positive bacteria with an outer membrane. All together, SCL data for more than 3 700 000 proteins are currently stored in cPSORTdb. This database is now set to be continually updated, with whole-proteome SCL predictions added as newly available sequenced prokaryotic genomes become available through NCBI's microbial genome database each month.

## 3.4 Automatic database update using a computational 'outer membrane detection' procedure

A major new feature of PSORTdb version 2 is the ability to automatically determine what 'Gram-stain' or cellular structure a given bacterial proteome should be analyzed under, by identifying through a sequence analysis whether the proteome is consistent with an outer membrane-containing bacterium or not. Previously, when microbial genomes have been released by NCBI, we manually determined the Gram-stain and cell structure based on bacterial phylum recorded and literature review. We noticed that neither the Gram-stain in the NCBI microbial database nor the bacterial phylum were 100% accurate in indicating the cell structure, due in part to the increasing diversity of bacterial genomes being sequenced. Gram-positive organisms traditionally have a cytoplasmic membrane surrounding the cell, and a thick cell wall composed of peptidoglycan that encircles the cytoplasmic membrane. Gram-negative organisms typically have a much thinner cell wall within the periplasm and an asymmetrical outer membrane surrounding the entire cell that is in addition to the cytoplasmic membrane. However, the traditional Gram-staining procedure does not always accurately denote the structure of all bacteria. For example, *Deinococcus* spp. stain Gram-positive because they have a thick cell wall, yet they also have an outer membrane (20). *Mycoplasma* spp., on the other hand, stain Gram-negative because they have no cell wall, but they also only have one cell membrane (i.e. no classical outer membrane) (21). The former should really be analyzed like a Gram-negative, to identify proteins in its outer membrane, while the latter should not have proteins predicted in non-existing outer membrane SCLs since it contains no such structure. Using taxonomy alone is also insufficient in detecting cell structure. Most bacteria within one phylum tend to have the same cellular structure, but *Halothermothrix orenii* of the phylum Clostridia has both characteristics of Gram-positive organisms yet also has an outer membrane (22). Hence, we developed a novel automatic cell-structure determination method, which we report here. Through research of different possibilities, we have determined that the presence of an outer membrane in a bacterium can be accurately determined by detecting the presence of the outer membrane protein Omp85, or more accurately the *omp85* gene or its orthologs, in a microbial genome. Omp85 is essential for outer membrane biogenesis and is the only known essential outer membrane protein for the viability of bacteria (the latter based on high-resolution analyses of saturated transposon mutagenesis of classic Gram-negative bacteria such as *Pseudomonas aeruginosa* (23,24). Using Omp85 proteins from four divergent genera of bacteria: *Neisseria gonorrhoeae*, *Thermosipho africanus*, *Synechococcus sp. PCC 7002* and *Thermus thermophilus*, we use BLAST to search for homologs of each in the sequenced bacterial genome to be analyzed (*E*-value cut off of $10^{-8}$). We found this was necessary to ensure high recall/sensitivity as simply using one Omp85 protein or ortholog did not detect all bacteria that we had manually confirmed as having an outer membrane. Using a data set of 813 diverse bacterial proteomes, curated regarding their outer membrane status, we also determined the appropriate *E*-value cutoff for this analysis and were able to easily obtain 100% precision and 100% recall using the diverse set of four Omp85 query sequences. We then compare the results with the phylum taxonomy of the bacteria, which are usually good indicators of bacterial membrane structure except in a few unusual cases. If the results from the two methods agree, the bacterial structural category is automatically assigned. If the two results disagree, then manual examination is used to assign one of the categories: classic Gram-negative, classic Gram-positive, Gram-positive with an outer membrane and Gram-negative without an outer membrane. Details of the analysis results can be found in Appendix B – supplementary data. In this way, the majority of the 1000+ prokaryotic genomes completely sequenced to date have been automatically assigned a cell structure and a corresponding PSORTb prediction module pipeline, with a few atypical bacteria flagged for manual inspection. The reason we require a combination of Omp85 BLAST and phylum analysis is that we find Omp85 analysis alone misses some organisms. For example, some species of *Buchnera* spp. did not have a positive result for Omp85. It is possible that these particular species actually do not possess an outer membrane, but laboratory examinations are required to confirm this. In addition, some genomes released seem to be incomplete or were of poor quality, in which case Omp85 was not identified from the BLAST analysis. This analysis not only aids appropriate automatic prediction of bacterial SCL prediction, but may also serve as a useful resource for microbiologists in general wishing to quickly determine the membrane structure of a given bacterial genome. Of course, there is the possibility of yet other atypical second membranes being discovered such as the unusual mycobacterial membrane (25), but manual curation of these cases should be possible and we do aim to increase the capability of the PSORTb family of software to identify such bacteria more accurately in the future.

## 3.5 Advanced search and filter functions

PSORTdb 2.0 is available for access with an updated web interface that has maintained the flexible search and browse functionalities, but with improved usability. One may search in either the database of proteins of known SCLs (ePSORTdb) or the database of computationally predicted SCLs (cPSORTdb) for specific proteins by organism domain, taxonomy based on NCBI's Taxonomy Database, Gram-stain, localization, secondary localization, protein name, GenBank ID number or a combination of any of these categories. There is an option to download the list of proteins that match the search criteria for the user's convenience. For example, one may obtain a list for all predicted extracellular proteins in all of the *Escherichia coli* strains by selecting 'Extracellular' for the Localization search category and '*Escherichia coli*' for Organism name search category.

If the user has the sequence of a protein that lacks an ID corresponding to PSORTdb's GenBank locus tag, our web interface provides a BLAST search function, which allows the user to find the localization of the query protein as well as homologous proteins. In the vast majority of the cases, SCL is highly conserved between highly similar, homologous proteins (26). In some cases proteins that are peripherally attached to the cell membrane or the cell wall will have only one of its two main localizations predicted depending on its sequence. These can be noted in the database with text ‘this protein may have multiple localizations’ or ‘unknown/multiple localization’ (the ‘unknown’ designation being because the scores for a given localization are split over multiple localizations, and we feel such cases should require more manual inspection to deduce their most probable localization/localizations). In some rare cases, proteins with homologous sequences have changed localizations (27). The BLAST option allows for the detection and examination of these cases.

To browse an entire replicon of a genome, the user can now start by simply entering an organism name. An auto-complete function allows the user to select from a list of possible organism/strain names once they start typing a name. Each genome replicon now has a proteome summary page. A pie chart of PSORTb 3.0 protein localization proportion distributions within the proteome, and numeric break down of the number of proteins in each localization category for both PSORTb 3.0 and PSORTb 2.0 predictions, and general taxonomy and cellular-structural information are provided (i.e. under what PSORTb analysis category the organism was analyzed, based on the ‘outer membrane detector’ analysis that had been performed). By clicking on a specific localization label, a list of all proteins from that localization within the organism proteome will be returned. This is useful if a researcher wants to get a list of, for example, all outer-membrane proteins in a pathogenic bacterium for drug or vaccine target research.

## 3.6 Conclusion

We have developed a new version of PSORTdb, which contains a greatly expanded data set of experimentally verified SCLs, as well as pre-computed highly precise SCL predictions for all prokaryotes, including now Archaea and bacterial organisms with atypical membrane/cell structures. The web server has been redesigned to facilitate user-friendly search and browsing of the database. It is continuously updated as newly sequenced prokaryotic genomes are released, using a novel computationally based cell-structure analysis which we developed. This allows PSORTdb to remain useful for researchers for analysis of novel species as the number of microbial genome sequences grow at a rapid pace. The contents of PSORTdb can be easily incorporated into whole-genome annotations and the entire database is open access, so it may be a valuable and convenient tool for a wide range of bioinformatics, genomics and microbiology researchers.

## 3.7 Funding

Natural Sciences and Engineering Research Council of Canada (NSERC) (with the support of the Cystic Fibrosis Foundation and Simon Fraser University Community Trust); Michael Smith Foundation for Health Research Senior Scholarship (to F.S.L.B.). Funding for open access charge: NSERC.

## 3.8 Acknowledgements

We thank NCBI for providing a compilation of complete microbial genomes data and gratefully acknowledge Raymond Lo for comments regarding this article.

## 3.9 References

1. Nakai, K. and Kanehisa, M. (1991) Expert system for predicting protein localization sites in gram-negative bacteria. Proteins, 11, 95-110.

2. Nakai, K. and Kanehisa, M. (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics, 4, 897-911.

3. Horton, P. and Nakai, K. (1997) Better prediction of protein cellular localization sites with the k nearest neighbors classifier. Proc. Int. Conf. Intell. Syst. Mol. Biol., 5, 147-152.

4. Bannai, H., Tamada, Y., Maruyama, O., Nakai, K. and Miyano, S. (2002) Extensive feature detection of N-terminal protein sorting signals. Bioinformatics, 18, 298-305.

5. Horton, P., Park, K.J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J. and Nakai, K. (2007) WoLF PSORT: Protein localization predictor. Nucleic Acids Res., 35, W585-7.

6. Gardy, J.L. (2003) PSORT-B: Improving protein subcellular localization prediction for gram-negative bacteria. Nucleic Acids Res., 31, 3613-3617.

7. Gardy, J.L., Laird, M.R., Chen, F., Rey, S., Walsh, C.J., Ester, M. and Brinkman, F.S. (2005) PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics, 21, 617-623.

8. Gardy, J.L. and Brinkman, F.S.L. (2006) Methods for predicting bacterial protein subcellular localization. Nat Rev Micro, 4, 741-751.

9. Yu, N.Y., Wagner, J.R., Laird, M.R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S.C., Ester, M., Foster, L.J., et al. (2010) PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics, 26, 1608-1615.

10. Lagesen, K., Ussery, D.W. and Wassenaar, T.M. (2010) Genome update: The 1000th genome--a cautionary tale. Microbiology, 156, 603-608.

11. Vallenet, D., Engelen, S., Mornico, D., Cruveiller, S., Fleury, L., Lajus, A., Rouy, Z., Roche, D., Salvignol, G., Scarpelli, C., et al. (2009) MicroScope: A platform for microbial genome annotation and comparative genomics. Database (Oxford), 2009, bap021.

12. Guo, T., Hua, S., Ji, X. and Sun, Z. (2004) DBSubLoc: Database of protein subcellular localization. Nucleic Acids Res., 32, D122-4.

13. Lu, P., Szafron, D., Greiner, R., Wishart, D.S., Fyshe, A., Pearcy, B., Poulin, B., Eisner, R., Ngo, D. and Lamb, N. (2005) PA-GOSUB: A searchable database of model organism protein sequences with their predicted gene ontology molecular function and subcellular localization. Nucleic Acids Res., 33, D147-53.

14. Hinz, U. and UniProt Consortium. (2010) From protein sequences to 3D-structures and beyond: The example of the UniProt knowledgebase. Cell Mol. Life Sci., 67, 1049-1064.

15. Zhou, M., Boekhorst, J., Francke, C. and Siezen, R.J. (2008) LocateP: Genome-scale subcellular-location predictor for bacterial proteins. BMC Bioinformatics, 9, 173.

16. Billion, A., Ghai, R., Chakraborty, T. and Hain, T. (2006) Augur--a computational pipeline for whole genome microbial surface protein prediction and classification. Bioinformatics, 22, 2819-2820.

17. Zhang, S., Xia, X., Shen, J., Zhou, Y. and Sun, Z. (2008) DBMLoc: A database of proteins with multiple subcellular localizations. BMC Bioinformatics, 9, 127.

18. Goudenege, D., Avner, S., Lucchetti-Miganeh, C. and Barloy-Hubler, F. (2010) CoBaltDB: Complete bacterial and archaeal orfeomes subcellular localization database and associated resources. BMC Microbiol., 10, 88.

19. Rey, S., Acab, M., Gardy, J.L., Laird, M.R., deFays, K., Lambert, C. and Brinkman, F.S. (2005) PSORTdb: A protein subcellular localization database for bacteria. Nucleic Acids Res., 33, D164-8.

20. Thompson, B.G. and Murray, R.G. (1981) Isolation and characterization of the plasma membrane and the outer membrane of Deinococcus radiodurans strain sark. Can. J. Microbiol., 27, 729-734.

21. Miyata, M. and Ogaki, H. (2006) Cytoskeleton of mollicutes. J. Mol. Microbiol. Biotechnol., 11, 256-264.

22. Mavromatis, K., Ivanova, N., Anderson, I., Lykidis, A., Hooper, S.D., Sun, H., Kunin, V., Lapidus, A., Hugenholtz, P., Patel, B., et al. (2009) Genome analysis of the anaerobic thermohalophilic bacterium Halothermothrix orenii. PLoS One, 4, e4192.

23. Voulhoux, R., Bos, M.P., Geurtsen, J., Mols, M. and Tommassen, J. (2003) Role of a highly conserved bacterial protein in outer membrane protein assembly. Science, 299, 262-265.

24. Tashiro, Y., Nomura, N., Nakao, R., Senpuku, H., Kariyama, R., Kumon, H., Kosono, S., Watanabe, H., Nakajima, T. and Uchiyama, H. (2008) Opr86 is essential for viability and is a potential candidate for a protective antigen against biofilm formation by Pseudomonas aeruginosa. J. Bacteriol., 190, 3969-3978.

25. Hoffmann, C., Leis, A., Niederweis, M., Plitzko, J.M. and Engelhardt, H. (2008) Disclosure of the mycobacterial outer membrane: Cryo-electron tomography and vitreous sections reveal the lipid bilayer structure. Proc. Natl. Acad. Sci. U. S. A., 105, 3963-3967.

26. Nair, R. and Rost, B. (2002) Sequence conserved for subcellular localization. Protein Sci., 11, 2836-2847.

27. Li, T., Huang, X., Zhou, R., Liu, Y., Li, B., Nomura, C. and Zhao, J. (2002) Differential expression and localization of Mn and Fe superoxide dismutases in the heterocystous cyanobacterium Anabaena sp. strain PCC 7120. J. Bacteriol., 184, 5096-5103.

# 4: Global analysis of protein subcellular localization trends in Bacteria and Archaea

***Manuscript in preparation.***

**Author list:** Nancy Y. Yu, Jennifer L. Gardy, Fiona S. L. Brinkman

**Author contribution:** NYY and FSLB co-designed the analyses and co-edited the manuscript. NYY performed all analyses in this study and wrote the manuscript. JLG performed an initial proportion analysis with a smaller set of bacterial genomes and wrote a preliminary draft of the manuscript.

## 4.1 Abstract

Recent methods for the prediction of protein subcellular localization (SCL) in bacteria have exceeded high-throughput laboratory methods for localization analysis in terms of both precision and recall. Highly accurate software now permits rapid and accurate identification of protein localization for large portions of deduced proteomes from sequenced genomes. The availability of such tools now enables the study of bacterial protein localization in a global, comparative genomics context. To date, no large-scale analysis has been published of proportion trends of all protein localizations among a wide variety of proteomes deduced from sequenced bacterial genomes, and archaeal protein SCL proportions have never been examined before.

We now report the first comprehensive study of protein SCL trends in relation to prokaryotic proteomes. We used PSORTb 3.0, the most precise prokaryotic SCL prediction software, to analyze 1188 sequenced bacterial and archaeal genomes to estimate the proportion of proteins localized to each cellular compartment. We found the number of proteins at each SCL to have strong linear correlations to deduced proteome size, and that the proportions of proteins at each localization site are markedly consistent across Archaea and Bacteria, reflecting groupings based on cellular structures of monoderms (encased by one membrane) and diderms (encased by two membranes). Our data implies that microbial evolution proceeds through the simultaneous acquisition or loss of protein sub-networks whose components span multiple localization sites. The findings highlight the need for developing bacterial group-specific SCL predictors based

on cellular structure more diverse than the classical Gram-positive and Gram-negative model. Additionally, the analysis suggests which localizations are most likely being under-predicted. Accurate and appropriate analysis of the diversity of microbial organisms is key from an applied microbiology perspective.

## 4.2 Introduction

Predicting the subcellular localization of bacterial proteins is an important first step in a number of processes, from genome annotation to drug/vaccine target identification. Knowing the localization of a protein can provide a clue to its function as well as identify its potential for industrial uses such as medical or environmental cell-surface biomarkers. Given its importance, many experimental studies have been performed in order to identify proteins that reside in subcellular compartments of various organisms [1-4]. Experimental approaches are costly, can only be used to on a small number of Archaea and Bacteria that can be cultured, and only subsets of genes are expressed at a given growth condition. Not all genes may be expressed, even if multiple conditions are tested. There is also the problem of contamination of proteins from other subcellular fractions in the process of protein fraction preparation, especially cross-contamination with abundant proteins [5]. Accurate computational prediction of protein subcellular localization can potentially overcome many of these limitations.

Examining localization globally for all proteins encoded within a genome is also useful. This may provide a clue to fundamental rules that bacteria follow regarding spatial organization of their protein interaction network. Comparing localization proportions across hundreds of genomes can potentially reveal important differences between species or groups of organisms, helping us understand their evolutionary history. Previously, it has been estimated that cytoplasmic membrane proteins account for approximately 20-30% of a given proteome (from diverse prokaryotic and eukaryotic organisms [6-8]. A number of computational SCL prediction tools have been developed since then with higher precision and recall [9-11]; however, only PSORTb can be easily applied to a large number of proteomes. Mass-spectrometry-based proteomic analysis of single organisms has been found to be in good agreement with tools such as PSORTb [1]. To our knowledge, no global studies have been performed on all SCLs of diverse groups of whole bacterial and archaeal deduced proteomes from complete genome sequences. At minimum, no such study has been performed at the level of accuracy now possible. With SCL prediction precision matching and even surpassing experimental results [5], with proteome prediction coverage of 81% for monoderms and 72% for diderms on average, at a precision of >95% for PSORTb, it is now possible to conduct such a study.

We therefore set out to examine protein localization proportion trends for deduced proteomes from over 1000 bacterial and archaeal complete genomes, using PSORTb 3.0, the most precise SCL prediction software [9]. We looked at commonalities among localization proportions among bacteria with one membrane (traditionally Gram-positive, now referred to as monoderms), two membranes (traditionally Gram-negative, now referred to as diderms), as well as Archaea. We also examined bacteria with specialized cellular structures, including monoderms with a Gram-negative phenotype (no cell wall) and diderms with different outer membranes. We examined trends in SCL proportions among strains of closely related bacteria of differing proteome size, which helps provide insight into how protein subcellular localization prediction may be potentially further improved to achieve yet higher accuracy for some proteomes that are less well predicted.

## 4.3 Results

### 4.3.1 Proportions of proteins at each SCL are consistent regardless of proteome size

The analysis examined global SCL trends for bacteria using the deduced proteomes from 370 monoderms and 724 diderms. Overall, the numbers of proteins for all localizations exhibit tight linear correlation to the deduced proteome size of both monodermic and didermic bacteria. Figure 4.1 shows the strongest linear correlations – the cytoplasmic SCL (r = 0.98 for monoderms; r = 0.97 for diderms), followed closely by cytoplasmic membrane SCL, with r = 0.97 for monoderms and diderms, respectively). These are the localizations for which there is the most data, since the majority of proteins in a cell are either cytoplasmic or associated with the cytoplasmic membrane. All other SCL vs. proteome size correlations can be found in Appendix C Supplementary data Figure 1. Using the trend line equations, the analyses suggest that, for monoderms, as the proteome increases in size, the number of proteins at each SCL increase at very consistent rates. Notably, the number of proteins predicted as "unknown" have much higher linear correlations with proteome size (r = 0.95 for monoderms and 0.94 for diderms) than the periplasmic, outer membrane, cell wall and extracellular localizations. There may be some sort of relationship between the ratios of protein SCLs and bacterial proteome sizes for particular bacteria. Deviation from linearity at small proteome sizes suggest that small proteomes in diderms contain more cytoplasmic proteins and fewer proteins in all other SCLs, including unknowns. Small proteomes in monoderms contain more cytoplasmic, cytoplasmic membrane and cell wall proteins, and fewer extracellular proteins and unknowns.

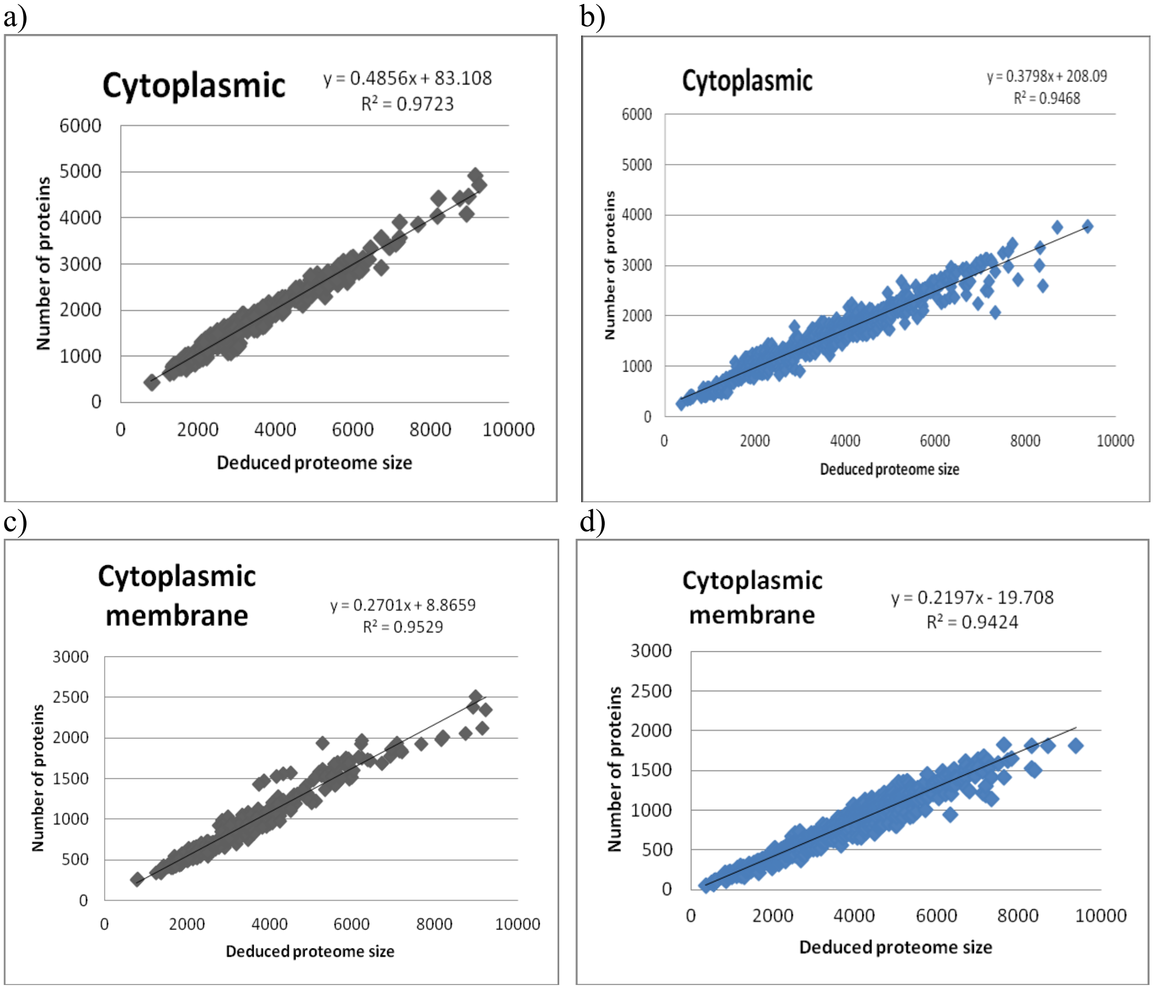
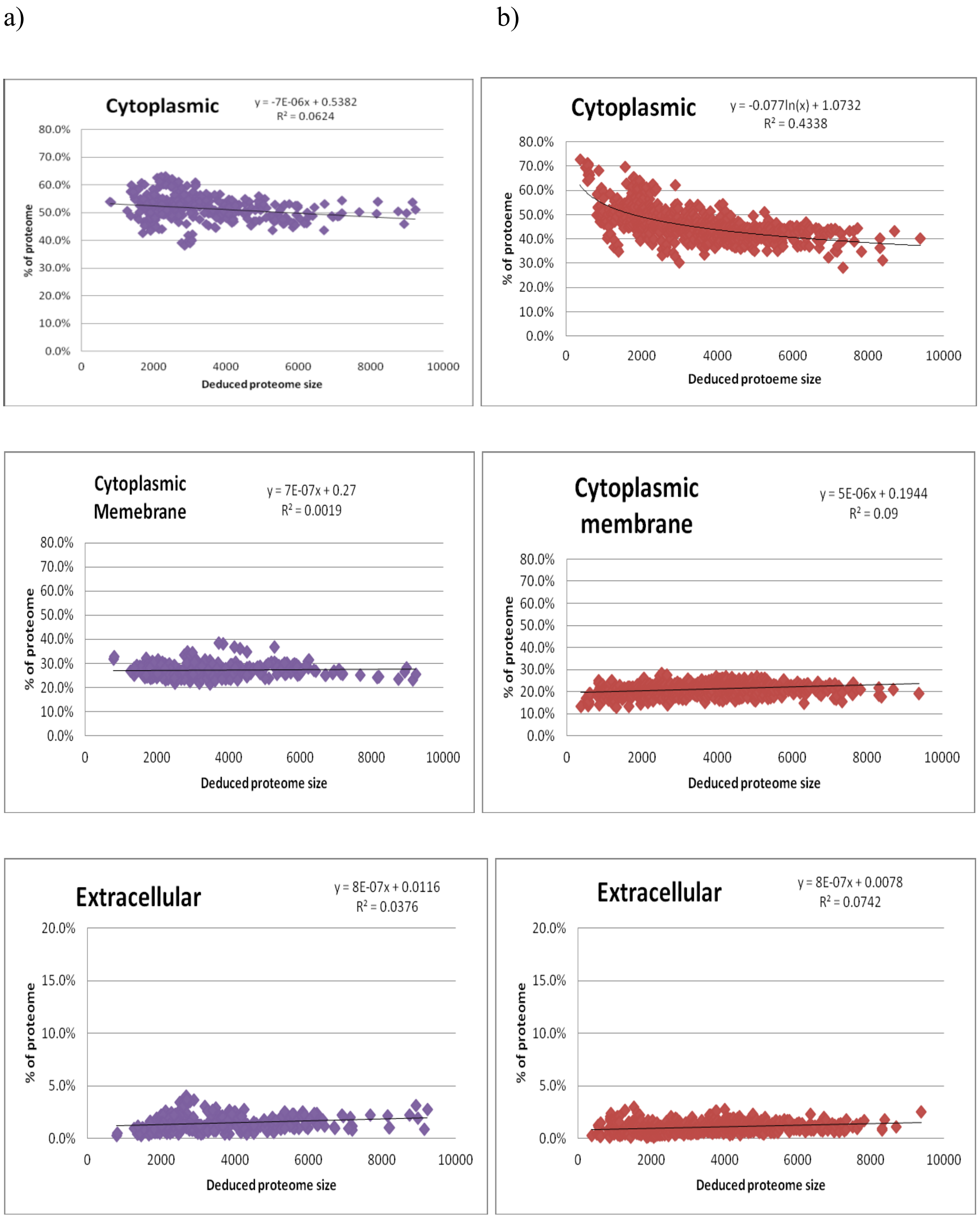


Figure 4.1 Number of cytoplasmic and cytoplasmic membrane SCLs have very high linear correlations for monodermic and didermic bacteria. a) monodermic cytoplasmic b) didermic cytoplasmic c) monodermic cytoplasmic membrane d) didermic cytoplasmic membrane

When we examine the data in terms of proportions of proteins at a given SCL, versus proteome size, proportions are remarkably consistent, regardless of proteome size. Figure 4.2 depicts a scatter plot of SCL proportions vs. proteome size for monoderms and diderms, with cytoplasmic membrane being the most consistent proportions of all SCLs. The slopes of the trend lines for the cytoplasmic membrane proportions are very small (in the scale of $10^{-7}$ for monoderms and $10^{-5}$ for diderms), with $R^2 < 0.1$ in both cases, which suggests that the proportion of cytoplasmic membrane SCLs is consistent, regardless of proteome size. The mean proportions of cytoplasmic membrane SCL are 27.2% and 21.0%, for monoderms, and diderms respectively, with a standard deviation of 2.6% for both groups, which is consistent with previous findings [7]. For monodermic bacteria, the percentage of cytoplasmic and cell wall SCLs seem to decrease slightly with increasing proteome size while extracellular and unknown SCL proportions appear to increase slightly as proteome size increases. However, these proportion increases and decreases are very small versus the corresponding change in proteome size occurring. The mean percentages of cytoplasmic, cytoplasmic membrane, cell wall, and extracellular proteins SCLs for monoderms are 51%, 27%, 1.1%, and 1.4%, respectively, with 18% of SCL predicted "unknown". For didermic bacteria, the percentage of cytoplasmic SCLs decreases relatively sharply as proteome size increases while all other SCLs increase slightly as proteome size increases. The average proportions of cytoplasmic, cytoplasmic membrane, periplasmic, outer membrane, and extracellular SCLs for diderms are 45%, 21%, 2.2%, 1.8%, and 1.0%, respectively. The proportion of didermic proteomes predicted as "unknown" averages at 28%, which is 9.3% higher than the less complicated monodermic proteomes. The slopes of the linear correlation trend lines for each SCL ranges from $10^{-7}$ to $10^{-5}$, with very low $R^2$ values for all SCL proportions ($R^2 < 0.1$) except for cytoplasmic ($R^2 = 0.42$) and periplasmic SCLs ($R^2 = 0.38$) for didermic bacteria. The most consistent appears to be cytoplasmic membrane SCL for monoderms, with the smallest slope of $7 \times 10^{-7}$ and $R^2 = 0$.


a)
b)
Cytoplasmic
y = -7E-06x + 0.5382
R² = 0.0624
% of proteome
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
0
2000
4000
6000
8000
10000
Deduced proteome size
Cytoplasmic
y = -0.077ln(x) + 1.0732
R² = 0.4338
% of proteome
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
0
2000
4000
6000
8000
10000
Deduced protoeme size
Cytoplasmic
Memebrane
y = 7E-07x + 0.27
R² = 0.0019
% of proteome
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
0
2000
4000
6000
8000
10000
Deduced proteome size
Cytoplasmic
membrane
y = 5E-06x + 0.1944
R² = 0.09
% of proteome
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
0
2000
4000
6000
8000
10000
Deduced proteome size
Extracellular
y = 8E-07x + 0.0116
R² = 0.0376
% of proteome
20.0%
15.0%
10.0%
5.0%
0.0%
0
2000
4000
6000
8000
10000
Deduced proteome size
Extracellular
y = 8E-07x + 0.0078
R² = 0.0742
% of proteome
20.0%
15.0%
10.0%
5.0%
0.0%
0
2000
4000
6000
8000
10000
Deduced proteome size

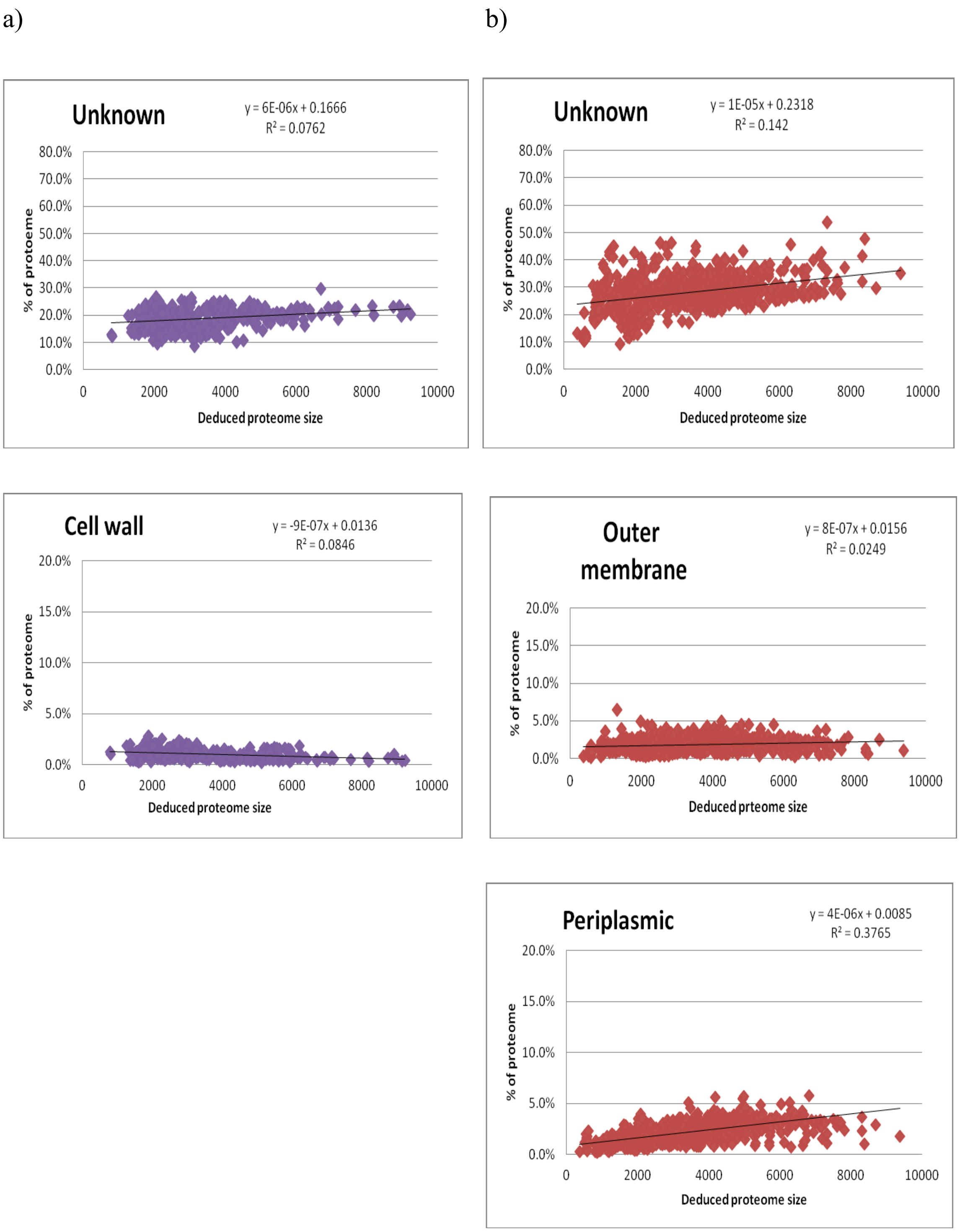

a)
b)
Unknown
y = 6E-06x + 0.1666
R² = 0.0762
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
% of protoeme
0
2000
4000
6000
8000
10000
Deduced proteome size
Unknown
y = 1E-05x + 0.2318
R² = 0.142
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
% of proteome
0
2000
4000
6000
8000
10000
Deduced proteome size
Cell wall
y = -9E-07x + 0.0136
R² = 0.0846
20.0%
15.0%
10.0%
5.0%
0.0%
% of proteome
0
2000
4000
6000
8000
10000
Deduced proteome size
Outer membrane
y = 8E-07x + 0.0156
R² = 0.0249
20.0%
15.0%
10.0%
5.0%
0.0%
% of proteome
0
2000
4000
6000
8000
10000
Deduced prteome size
Periplasmic
y = 4E-06x + 0.0085
R² = 0.3765
20.0%
15.0%
10.0%
5.0%
0.0%
% of proteome
0
2000
4000
6000
8000
10000
Deduced proteome size

Figure 4.2 Proportion of protein SCLs for a) monoderms and b) diderms. Cytoplasmic membrane SCL is the most consistent.

Despite a strong linear correlation between the number of cytoplasmic proteins and bacterial proteome size, it appears that the proportion of predicted cytoplasmic protein slightly decreases with increasing proteome size. Notably, the proportion of unknowns is inversely correlated to the proportions of cytoplasmic SCLs both for monoderms (r = -0.80) and for diderms (r = -0.85). No significant correlations were found between unknowns and other SCLs.

### 4.3.2 Analysis among some phylogenetic groups showed correlation between some SCL proportions

We wanted to analyze if there are more observable proportion trends among phylogenetically related groups of bacteria which are highly similar to each other yet differ notably in proteome size. The genome sequences of 30 strains of *Escherichia coli* were available, with deduced proteome sizes ranging from 4084 to 5516 predicted proteins. Figure 4.3 shows a scatter plot of SCL proportions in relation to proteome size for 30 *E. coli* strains. We looked at the trends of SCL proportions and found that the percentage of cytoplasmic, cytoplasmic membrane and periplasmic SCLs are positively correlated (r > 0.70). As a group, these SCL proportions are negatively correlated to total proteome sizes, the proportion of unknowns, and the proportion of extracellular SCLs. We examined *Pseudomonas spp*. as well (organisms noted for their relatively large genome size and metabolic capacity), versus *E. coli* which are noted for phage insertions causing a large proportion of genome size variation. SCL proportions were very consistent for the *Pseudomonas* proteomes (Figure 4.4). However, in both cases, the cytoplasmic SCLs are strongly inversely correlated to the “unknown” SCLs (r= -0.96 for E. coli strains and r = -0.85 for *Pseudomonas* spp.), in keeping with the larger studies of more diverse bacteria.

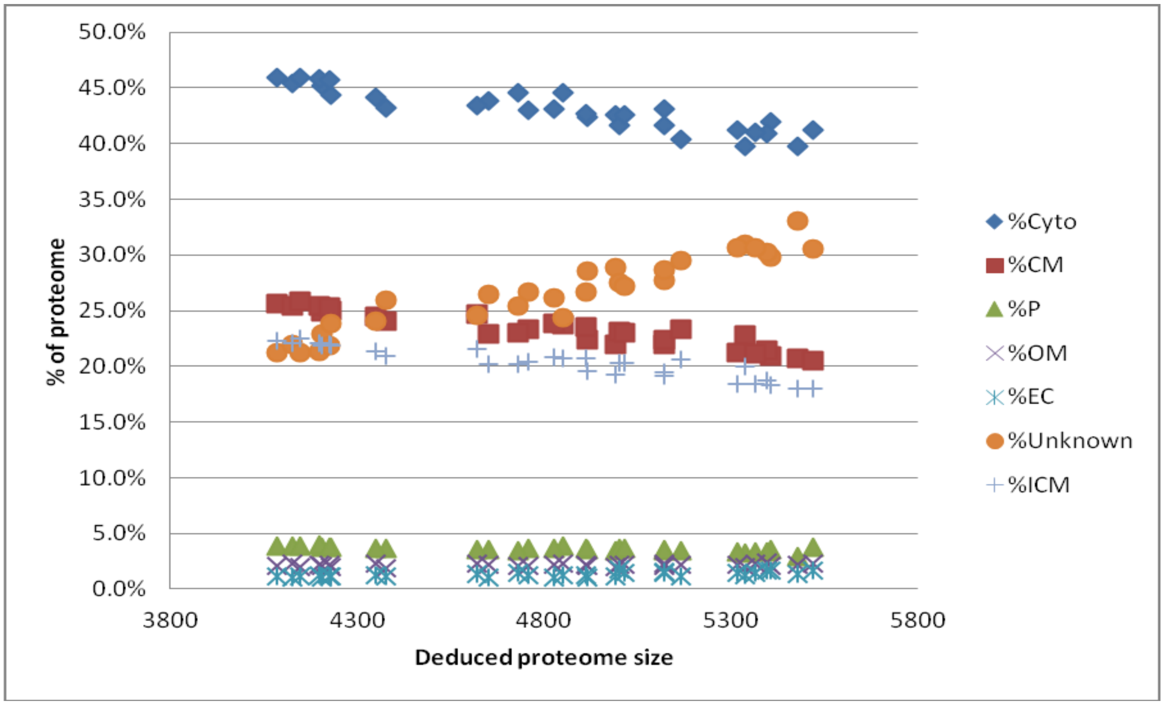


Figure 4.3 Proportions of all SCLs for 30 *Escherichia coli* strains. Correlations between SCLs are more apparent.

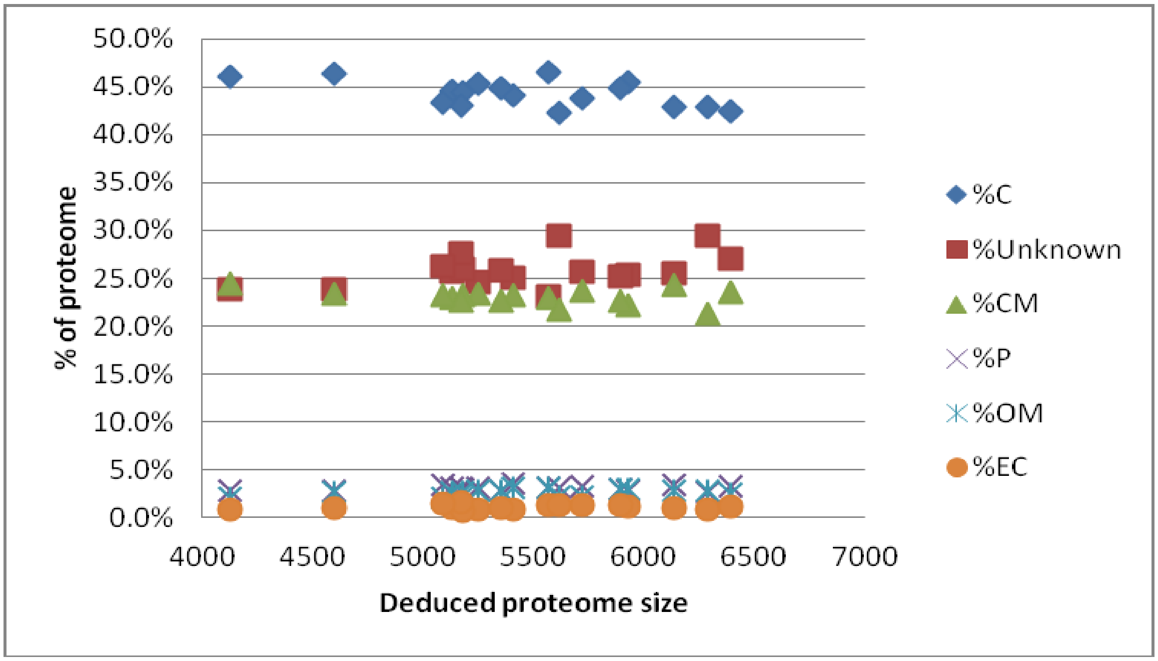


Figure 4.4 Proportions of all SCLs for 17 *Pseudomonas* species.

### 4.3.3 Bacteria with ‘atypical’ cellular structures can be detected from proportion analysis

By visually inspecting the scatter plot of proportions of cytoplasmic membrane SCL vs. deduced proteome size for monoderms (Figure 4.2), it is apparent that a few data points for monoderm cytoplasmic membrane proportions are higher than all other data points. Upon further examination, these points belong to organisms of the phylum Chloroflexi, a group of monodermic bacteria that contain chlorosomes, which are membranous organelles found within the cells of members of this phylum [12]. This is an interesting case demonstrating that, while most bacteria sequenced so far follow a general proportion trends, organisms with ‘atypical’ cellular structures deviating from the trend can be detected just from an SCL proportion analysis (note: ‘atypical’, at least in the context of what prokaryotic genomes have been sequenced to date). In this case, proteins destined for chlorosome membranes are predicted as cytoplasmic membrane SCL by PSORTb. However, it suggests that with an SCL proportion analysis, other organisms with especially low or high proportions of proteins at a given SCL could also contain special organelles or cellular structures in general can be detected as well. For example, *Buchnera* spp. (with proteome sizes under 580 proteins) appear to have a higher proportion of cytoplasmic proteins and lower proportions of other SCLs predicted. The deviation from the average SCL proportions might reflect the unique nature of such obligate endosymbionts living within eukaryotic cells (a nutrient-rich environment), which require less cell envelope proteins and secreted proteins for nutrient transportation and stress-response detection.

### 4.3.4 Bacteria with atypical cellular structures tend to have lower proteome prediction coverage

Besides organisms of the Chloroflexi phylum, there are other bacterial organisms that are known to have atypical cellular structures. Classically, Gram-positive organisms are monoderms and Gram-negative bacteria are diderms. Table 4.1 provides a list of bacterial organisms that either do not conform to the classical Gram-stain identification of number of cell membranes, or harbor atypical subcellular structures. The Genus *Deinococcus* spp. stain Gram-positive, however they also have an outer membrane [13]. Organisms belonging to the phylum of Tenericutes stain Gram-negative because they have no cell wall, however they also do not have an outer membrane [14]. *Mycobacterium* spp. have a unique outer membrane that is not likely of the same origin as other didermic outer membranes [15]. *Rhodopirellula baltica* and *Cyanobacteria* spp have additional membrane bound structures within the cell [16,17]. These atypical organisms all seem to be less well-predicted by PSORTb 3.0 in terms of proteome coverage compared to "typical" or model organisms.

Table 4.1 Examples of bacteria that either harbor atypical subcellular structures, or do not conform to the classical Gram-stain cell structure, in particular in terms of the identification of the number of cell membranes.

| **Bacterial organisms** | **Gram-stain** | **# of membranes** | **Has a Cell wall** | **Other structures** | **Comments on SCL predictions** |
|---|---|---|---|---|---|
| *Deinococcus* spp. | Positive | 2 | Thick | | Lower than average proteome coverage (67%) |
| Chloroflexi class | Positive | 1+ | Thick | Chlorosome Membrane stuctures | %CM much higher than all other monoderms (35%) |
| *Mycobacterium* spp. | Positive | 1+ | Thick | Mycomembrane | Cell wall/cell surface proteins |
| *Cyanobacteria* spp. | Negative | 2+ | Thin | Thylakoid membrane | Lower than average overall coverage (64%) |
| *Rhodopirellula baltica* | Negative | 2+ | Thin | Membrane-bound organelles | Really poor overall coverage (46%) |
| Tenericutes | Negative | 1 | No | | Low overall coverage compared to monoderms (68%) |

### 4.3.5 Archaea SCL trends are comparable to Bacteria, with higher proportion of predicted cytoplasmic proteins and lower cytoplasmic membrane proteins

We also analyzed 94 archaeal proteomes to compare proportions for this fundamentally different domain of life. The proteome prediction coverage of Archaea is surprisingly high, given that the archaeal predictors were trained on mostly bacterial proteins with a smaller dataset of archaea-specific proteins. Figure 4.5 shows a summary scatter plot of all archaeal SCL proportions vs. deduced proteome size. The average proportion of cytoplasmic SCL is much higher than bacteria (67% vs. 51% for monodermic bacteria; statistically significantly different $p < 0.05$). The proportion of cytoplasmic membrane proteins (18%) was lower than monodermic and even didermic bacteria (21%). Even with the higher cytoplasmic proportion predictions, the proportion of Unknowns still appears to be correlated to the cytoplasmic proportions ($r = -0.78$).

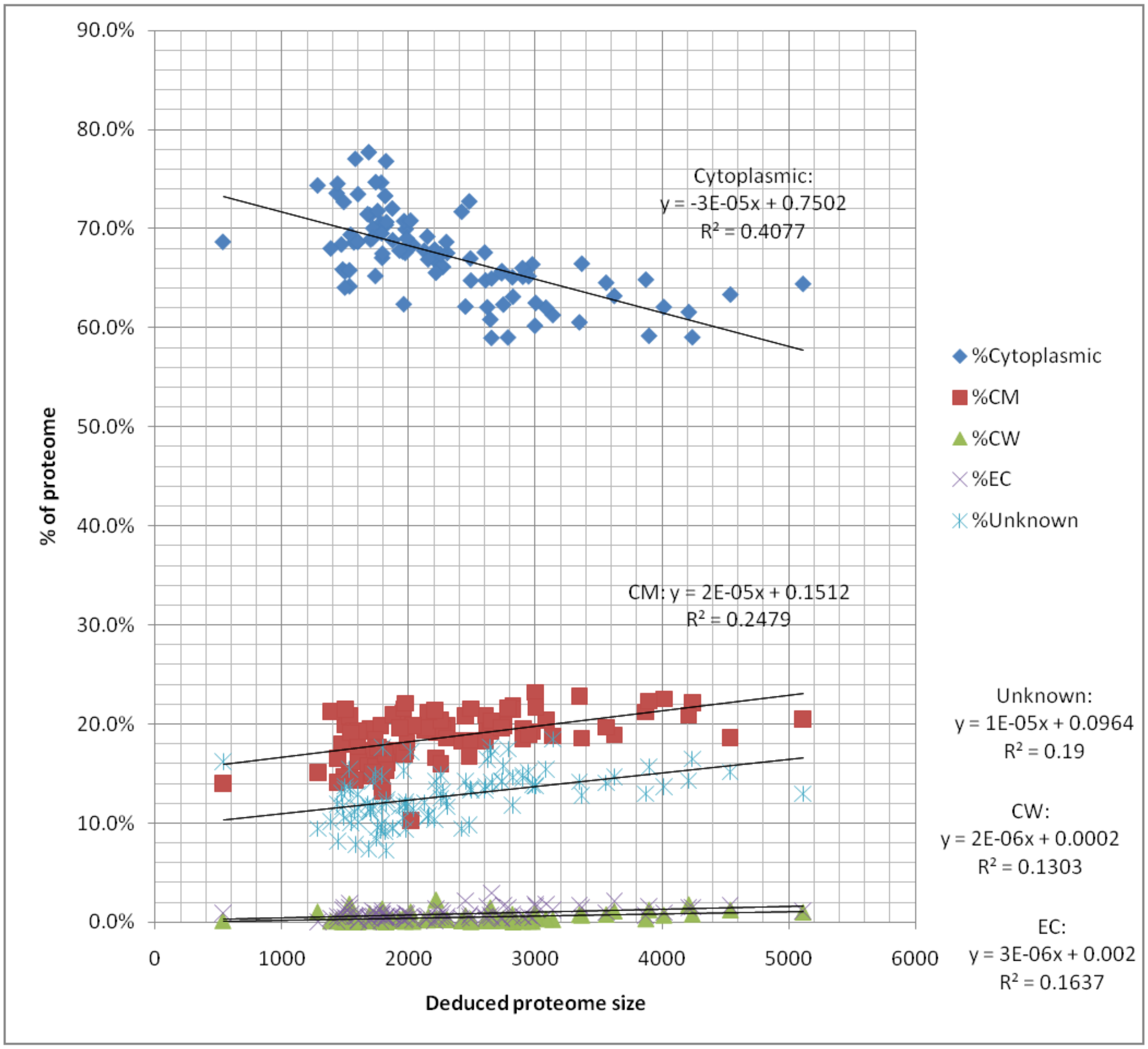


Figure 4.5 Proportions of all SCLs for archaeal proteomes are similar to bacterial SCL proportions, with more cytoplasmic SCLs and fewer membrane SCLs predicted. Cell wall and extracellular SCLs are inconsistent, likely due to the fact that prediction sensitivity and specificity are lower than the other SCL predictions.

## 4.4 Discussion

This is the first study examining trends across all SCL proportions for 1188 diverse Bacteria and Archaea whose genomes were sequenced at the time of writing. Most strikingly, despite diverse cell shapes, proteome sizes, habitats, and lifestyle of bacterial organisms, the proportions of proteins at a given SCLs are remarkably consistent for microbes with similar cell/membrane structures. The data suggests that, as proteomes increase in size, protein sub-networks that traverse membrane boundaries are added to the bacterial and archaeal cell. Presumably, these most likely reflect sub-networks such as transport systems taking up different substrates or efflux compounds (such as toxic compounds like antimicrobials) from the cell [18, 19]. It has been observed that bacteria that are more free-living in soil and water tend to have larger proteomes [20], presumably to be able to better adapt to their wildly changing environmental conditions and take up a wider range of substrates as sources of nutrients. Our analysis suggests that such transport systems have a common structure, traversing membrane boundaries with proteins at proportional numbers for the different localizations and provide insight into cellular sub-network structure. Presumably there must be multiple cytoplasmic proteins to process a given new metabolic capability added by a sub-network, coupled with some membrane proteins, and selected additional secreted proteins in some cases, explaining the proportions observed. For example, it has been previously noted that transcription factors seem to increase at roughly quadratic rate with proteome size [18,19]. Furthermore, it is generally noted that genes in certain high-level categories, notably metabolic genes, scale as a power-law proportional to the genome size [21,22]. Our analysis, the most accurate of its kind currently possible, now indicates that protein

SCLs scale with proteome size as well. Many differences in these consistencies in proportions, for those organisms with the same cellular structure/sub-structures, can be attributed to problems with not predicting certain protein localizations as well as others. For example, the prediction of extracellular, cell wall, and periplasmic protein SCLs have lower accuracy compared to membrane and cytoplasmic SCLs. Also there appears to be difficulty with predicting proteins not as well annotated such as proteins, including phage proteins, found in larger proteomes in some cases. It is well known that phage proteins are disproportionately poorly annotated, due to the larger gene pool associated with phage (increasing the chances of seeing something "new") [23]. The more marked inverse correlation between cytoplasmic proportions and unknown proportions in *E. coli*, with the proportion of unknowns increasing as proteome size increased, may reflect the large number of phage genes in larger *E. coli* genomes, but this requires further study.

Because of this strong inverse correlation between protein SCLs predicted as "unknown" and cytoplasmic SCLs, we suspect that most unknowns are probably cytoplasmic, with a much smaller proportion being of the other localizations. Cytoplasmic proteins are diverse, have no signal peptides, and do not have intrinsic motifs that make them cytoplasmic as opposed to non-cytoplasmic. They do not have structural motifs that anchor them at a particular localization, such as alpha helices for cytoplasmic membrane proteins and beta barrels for outer membrane proteins; cytoplasmic proteins that have no sequence similarity to the training dataset or are not picked up by the SVM cannot be identified by SCL predictors. Given that cytoplasmic proteins do not have signals to process them to a given localization, improved localization prediction may have to increasingly depend on increased training data based on proteins of experimentally determined localization, coupled with a sequence similarity-based identification of localization. Since localization is so conserved across species, this kind of approach is quite feasible [24]. Further developments in protein feature identification and machine learning algorithms may also help improve SCL prediction. Another hard to predict SCL is peripheral membrane proteins – proteins anchored to the cell wall or the inner or outer membrane by a peptide, a lipid moiety, or via some other molecule, but functional domain is localized within the cell, in the periplasm, or on the cell surface. Such proteins can have properties reflective of the multiple localizations that may only be partially picked up by the appropriate SCL prediction modules. Improvements in the sensitivity of all SCL prediction modules and a better decision tree (e.g. determining whether a protein is cytoplasmic or cytoplasmic membrane SCL, or both) are keys to making a better SCL predictor.

It is noted that model organisms such as *Escherichia coli* had higher than average prediction coverage, which is expected, since a good proportion of the training dataset comes from *E. coli* proteins. In general, smaller proteomes received higher prediction coverage than larger proteomes, since these larger proteomes likely contain genomic islands with novel genes not related to prediction software's known set of SCLs [23]. It is also apparent that bacterial organisms with atypical structures as well as organisms from phyla that are less related to the model organisms have lower than average proteome prediction coverage. Advances in prediction algorithms and homology searches may not significantly improve the prediction coverage for these organisms. Alternative strategies are needed, such as species-specific proteomic studies that will generate high quality dataset of experimentally determined SCL for these organisms, and then use this data for building better SCL predictors for these more diverse species. Bioinformaticians researching computational SCL prediction should also recognize that more bacteria are being sequenced with non-classical structures. For example: there are extra membrane structures for Chloroflexi, a lack of cell walls for Tenericutes, and the mycobacterial-specific outer membrane for *Mycobacterium* spp. Instead of trying to predict the classical monoderm SCLs for these organisms, future SCL prediction software should better take these specialized structures into account. Some development of phyla specific predictors has occurred, such as TBpred for *Mycobacterium* spp. [25]. However, there are many diverse, 'atypical' other bacterial species for which more specific predictors need to be developed (or at least a program such as PSORTb needs to be modified in the future to enable prediction based on more options for selecting more diverse phyla). Such prediction of more diverse species, particularly those currently poorly understood, may only be possible with the generation of high quality experimentally species-specific protein SCL data as training and testing datasets, coupled with slight changes in algorithms to accommodate different cellular structures.

Given the fundamental difference between archaeal and bacterial proteome content, it is surprising that Archaea proteomes prediction coverage is so high, even though the PSORTb prediction module was built predominantly with bacterial training dataset. It is possible that the prediction precision for Archaea is lower and some of these might be false positives. There is a need to generate accurate experimental SCL data to verify the prediction results however certainly initial examinations of the accuracy of the archaeal SCL prediction for PSORTb version 3.0 suggested that a high rate of false positives was not occurring [9]. The cytoplasmic membrane predictions should at minimum be quite accurate, since the training data for the alpha-helix transmembrane predictor contains many archaeal proteins. What is known from analyzing archaeal gene contents is that archaeal genomes tend to contain less carbohydrate metabolism and cell membrane genes, and more genes in the categories of energy production, coenzyme metabolism, and "unknown function" [20]. Structurally, its membrane lipids are ether-linked rather than ester-linked as in Bacteria and Eukarya [26]; cell walls in the domain of Archaea can consist of polysaccharides, protein polymers, or pseudo-peptidoglycan rather than peptidoglycan-based cell walls in Bacteria [27]. Archaea also do not seem to possess homologs to the typical specialized bacterial secretion systems [28]. For these reasons, the cell wall and extracellular protein SCLs may be under-predicted and not picked up by the prediction software. However, a high rate of cytoplasmic SCL identification makes it unlikely that Archaea contains higher proportions of cytoplasmic membrane, cell wall and extracellular proteins than bacteria. In this study we analyzed all Archaea as one group, but in reality the phylogeny and cellular structure is very diverse [29]. A more refined SCL predictor may better predict the cell wall and extracellular proteins for different types of Archaea.

In conclusion, our global analysis of protein subcellular localizations, predicted for over 1000 bacterial and archaeal proteomes, suggests that there is a typical sub-network structure repetitively added to a cell proteome as the proteome size increases, providing insight into protein network complexity. The results suggest further that many of the unknown predictions are likely cytoplasmic (and to some extend other localizations, as dictated by trend lines for SCL proportions as a function of proteome size). This analysis also highlights bacterial structures that are more complex than simple ‘classic’ monoderms and diderms. New approaches are be needed for improving SCL prediction for bacteria that deviate from non-classic cellular structures. Characterizing the SCLs of the remaining unpredicted proteins in selected organisms from certain key lineages, reflecting different subcellular structural arrangements, could provide further key insights into the evolution of cell networks as well as having the applied benefit of potentially aiding drug development, vaccine design and diagnostic target discovery.

## 4.5 Materials and Methods

### 4.5.1 Dataset

Deduced proteomes of completely sequenced bacterial and archaeal genomes were downloaded from NCBI’s FTP site October 2010 (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Genomes were sorted into monodermic bacteria, didermic bacteria, or Archaea using a combination of domain/phylum identification as well as the presence or absence of Omp85, an essential outer membrane protein, as described in detail in [30]. Organisms with structures not conforming to classic Gram-stain identification were not included in the SCL proportions analysis and are part of the list of organisms with atypical cellular structures in Table 4.1. Proteomes with fewer than 300 proteins were excluded from the analysis. Genomes with problems such as partially incompletely sequenced/annotated replicons were also excluded from the study. 370 monoderms, 724 diderms, and 94 archaeal proteomes were used to compute SCL proportions.

### 4.5.2 SCL calculations

For each organism, PSORTb 3.0 [9] was used to compute protein SCLs for the entire deduced proteome, using settings appropriate for the organism's domain/cell structure. The number of proteins at each protein SCL was plotted against deduced proteome size for that genome. For monoderms and Archaea, the SCLs predicted are cytoplasmic, cytoplasmic membrane, cell wall, extracellular, and "unknown" (prediction confidence below threshold). For diderms, the SCLs predicted include cytoplasmic, cytoplasmic membrane, periplasmic, outer membrane, extracellular, and unknown. For each domain and bacterial structural category, a scatter plot was made for the number of proteins versus proteome size for each SCL. The percentage of proteins at each SCL with regard to the whole proteome was computed. For each organism, the total proteome included all replicons (chromosomes and plasmids). We have also done the analysis using primary chromosomes only. Since the resulting proportion trends did not differ significantly we only showed the results of whole proteome analysis.

## 4.6 Acknowledgements

The authors wish to thank Carolyn Huston, Matthew Whiteside and Erin Gill of Simon Fraser University for valuable discussions. This work was funded by Natural Sciences and Engineering Research Council of Canada. FSLB is a Michael Smith Foundation for Health Research Senior Scholar and a Canadian Institutes of Health Research New Investigator.

## 4.7 References

1. Callister SJ, Dominguez MA, Nicora CD, Zeng X, Tavano CL, et al. (2006) Application of the accurate mass and time tag approach to the proteome analysis of sub-cellular fractions obtained from rhodobacter sphaeroides 2.4.1. aerobic and photosynthetic cell cultures. J Proteome Res 5(8): 1940-1947.

2. Viratyosin W, Ingsriswang S, Pacharawongsakda E, Palittapongarnpim P. (2008) Genome-wide subcellular localization of putative outer membrane and extracellular proteins in leptospira interrogans serovar lai genome using bioinformatics approaches. BMC Genomics 9: 181.

3. Rajalahti T, Huang F, Klement MR, Pisareva T, Edman M, et al. (2007) Proteins in different synechocystis compartments have distinguishing N-terminal features: A combined proteomics and multivariate sequence analysis. J Proteome Res 6(7): 2420-2434.

4. Fulda S, Huang F, Nilsson F, Hagemann M, Norling B. (2000) Proteomics of synechocystis sp. strain PCC 6803. identification of periplasmic proteins in cells grown at low and high salt concentrations. Eur J Biochem 267(19): 5900-5907.

5. Rey S, Gardy JL, Brinkman FS. (2005) Assessing the precision of high-throughput computational and laboratory approaches for the genome-wide identification of protein subcellular localization in bacteria. BMC Genomics 6: 162.

6. Wallin E, von Heijne G. (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Sci 7(4): 1029-1038.

7. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. (2001) Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes. J Mol Biol 305(3): 567-580.

8. Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, et al. (2005) PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics 21(5): 617-623.

9. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, et al. (2010) PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 26(13): 1608-1615.

10. Lu Z, Szafron D, Greiner R, Lu P, Wishart DS, et al. (2004) Predicting subcellular localization of proteins using machine-learned classifiers. Bioinformatics 20(4): 547-556.

11. Zhou M, Boekhorst J, Francke C, Siezen RJ. (2008) LocateP: Genome-scale subcellular-location predictor for bacterial proteins. BMC Bioinformatics 9: 173.

12. Sutcliffe IC. (2010) A phylum level perspective on bacterial cell envelope architecture. Trends Microbiol 18(10): 464-470.

13. Thompson BG, Murray RG. (1981) Isolation and characterization of the plasma membrane and the outer membrane of deinococcus radiodurans strain sark. Can J Microbiol 27(7): 729-734.

14. Miyata M, Ogaki H. (2006) Cytoskeleton of mollicutes. J Mol Microbiol Biotechnol 11(3-5): 256-264.

15. Niederweis M, Danilchanka O, Huff J, Hoffmann C, Engelhardt H. (2010) Mycobacterial outer membranes: In search of proteins. Trends Microbiol 18(3): 109-116.

16. Fuerst JA, Sagulenko E. (2011) Beyond the bacterium: Planctomycetes challenge our concepts of microbial structure and function. Nat Rev Microbiol 9(6): 403-413.

17. Nickelsen J, Rengstl B, Stengel A, Schottkowski M, Soll J, et al. (2011) Biogenesis of the cyanobacterial thylakoid membrane system--an update. FEMS Microbiol Lett 315(1): 1-5.

18. Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrener P, et al. (2000) Complete genome sequence of pseudomonas aeruginosa PAO1, an opportunistic pathogen. Nature 406(6799): 959-964.

19. McLeod MP, Warren RL, Hsiao WW, Araki N, Myhre M, et al. (2006) The complete genome of rhodococcus sp. RHA1 provides insights into a catabolic powerhouse. Proc Natl Acad Sci U S A 103(42): 15582-15587.

20. Konstantinidis KT, Tiedje JM. (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. Proc Natl Acad Sci U S A 101(9): 3160-3165.

21. van Nimwegen E. (2003) Scaling laws in the functional content of genomes. Trends Genet 19(9): 479-484.

22. Molina N, van Nimwegen E. (2009) Scaling laws in functional genome content across prokaryotic clades and lifestyles. Trends Genet 25(6): 243-247.

23. Nair R, Rost B. (2002) Sequence conserved for subcellular localization. Protein Sci 11: 2836-2847.

24. Hsiao WW, Ung K, Aeschliman D, Bryan J, Finlay BB, et al. (2005) Evidence of a large novel gene pool associated with prokaryotic genomic islands. PLoS Genet 1(5): e62.

25. Rashid M, Saha S, Raghava GP. (2007) Support vector machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. BMC Bioinformatics 8: 337.

26. White D. (2000) In: Anonymous The Physiology and Biochemistry of Prokaryotes. Oxford: Oxford University Press. pp. 565.

27. Cavicchioli R. (2007) In: Anonymous Archaea: Molecular and Cellular Biology. : ASM Press. pp. 450.

28. Eichler J. (2000) Archaeal protein translocation crossing membranes in the third domain of life. Eur J Biochem 267(12): 3402-3412.

29. Brochier-Armanet C, Forterre P, Gribaldo S. (2011) Phylogeny and evolution of the archaea: One hundred genomes later. Curr Opin Microbiol 14(3): 274-281.

30. Yu NY, Laird MR, Spencer C, Brinkman FS. (2011) PSORTdb--an expanded, auto-updated, user-friendly protein subcellular localization database for bacteria and archaea. Nucleic Acids Res 39(Database issue): D241-4.

# 5: Discussion and final remarks

For my thesis project, I have developed a new version of PSORTb that makes highly precise SCL predictions (>95% precision) with high recall/coverage for most bacterial proteomes and the first predictions specifically targeted for Archaea proteome analysis. I have developed the most comprehensive protein SCL database for prokaryotes, with a new capability for monthly auto-updating as newly sequenced bacterial and archaeal genomes become available. This auto-updating is based on a novel method I developed to predict whether a bacterium was likely a monoderm or diderm. Finally, I have performed the most comprehensive SCL proportion analysis for prokaryotes and established that prokaryotic genomes seem to evolve following a surprising universal SCL proportion rule. There are some exceptions, especially for organisms with more 'atypical' cell/membrane structures, but in general, the proportions of proteins at each SCL for all prokaryotic proteomes seem to be proportional to proteome size.

I wish to discuss a few final issues in this final chapter. First of all, while reviewing SCL prediction software published over the past 5 years, I noticed that quite a few web servers for software published earlier are no longer active or maintained. Many of the SCL databases have also either disappeared or have not been updated since the paper publication. Unlike algorithm development or biological discovery, where the results and experimental methods are useful own their own as published literature, SCL prediction software (or bioinformatics software and databases in general) are highly useful resources that benefit from being continually available and maintained (either as web resources, or standalone software maintained to be compatible with current operating systems) to assist biologists and bioinformaticians in accleerating their research. While new SCL prediction software manuscripts are frequently published, most seem to be proof-of-concept software, which do not offer use-friendly interface for batch protein sequence submission, nor do they provide a version of software that allows for easy integration into a bioinformatics pipeline. Many SCL prediction software developers also insist on optimizing the "accuracy" values, which strikes a balance between precision and recall, which is very different from PSORTb's long-term research goal of maintaining high precision, and working on gradually improving recall. Because of this, it is difficult to compare these algorithms to PSORTb in a useful way, or decide whether new algorithm variations would be worth being adapted into future generations of PSORTb. So far, the only other software that complements PSORTb well and seems to be of comparable precision, with still high recall, is Proteome Analyst. A combination of the two approaches would maximize proteome prediction coverage while maintaining high precision. However, this approach was not performed for the global proteome analysis because Proteome Analyst software is no longer available.

While working on this project, I discovered that many bacteria with more unusual cellular structures have lower prediction coverage than bacteria that are closer to the "classical" monodermic and didermic cell structures. The gene contents of these organisms also seem to be more divergent (less similar to sequences currently in gene databases), which is why the SCLs of a large proportion of these proteins are not predicted using the current training set of protein features. Simply researching general algorithm improvements may not be sufficient in improving predictions for these organisms. Future SCL prediction research should incorporate more experimentally based protein SCL data, as well as algorithm improvement targeted for less studied species.

With metagenomics becoming more commonplace, the next SCL prediction method should take into account making predictions from partial DNA gene and protein sequences. With the ability to sequence bits of DNA in the environment for unculturable bacterial and archaeal organisms, we are bound to come across many genes encoding proteins that have very low sequence similarity to annotated genes in current databanks. The ability to accurately predict the SCLs of these novel genes will be crucial for providing some clues to their potential functions. Currently no SCL predictor has yet been developed that is suitable for metagenomics datasets.

The field of SCL prediction for bacteria has come a long way, now with an average of 72% coverage for didermic bacteria and 81% for monodermic bacteria, at >95% precision. By improving algorithms, combining predictors, and incorporating key experimental SCL predictions for non-classical proteins into training data, it may be that about 90-95% of many bacterial and archaeal proteomes can be assigned with accurate SCLs. Future efforts will need to focus on species-specific proteins, especially for organisms with atypical structures and whose proteome prediction coverage seems much lower than average. It would also be useful to name the secretion pathways that deliver each of the exported proteins; however, this will only be possible with better understanding of the secretion pathways and with bigger collections of experimentally identified proteins that are exported by each secretion system to reach their destined localizations. Hopefully, all the efforts in improving prediction recall will also benefit

SCL prediction of metagenomics data as well, which is key since proteomics methods could prove particularly difficult for the analysis of such large collections of (predominantly unculturable) microbes. Computational protein SCL prediction is already exceeding the accuracy of certain high-throughput laboratory proteomics methods for SCL determination, and this work described here has improved it further. However, clearly, as we learn more about the diversity of microbes on the planet, SCL prediction should be further improved to better make predictions that factor in this incredible microbial diversity.

# APPENDICES

# Appendix A - Supplementary data for chapter 2

**Table 1: Performance comparison for Gram-positive SCL prediction software**

| Software† | Precision†† | Recall†† | Accuracy†† | MCC†† |
|---|---|---|---|---|
| **PSORTb 3.0** | | | | |
| C | 99.0 | 96.9 | 96.8 | 0.91 |
| CM | 96.0 | 86.2 | 97.7 | 0.90 |
| CW | 88.9 | 72.7 | 99.6 | 0.80 |
| EC | 92.9 | 67.6 | 97.3 | 0.78 |
| Overall | ***98.2*** | ***93.1*** | ***97.9*** | ***0.79*** |
| **PSORTb 2.0** | | | | |
| C | 98.5 | 93.9 | 94.1 | 0.84 |
| CM | 99.0 | 79.4 | 97.2 | 0.87 |
| CW | 92.9 | 59.1 | 99.5 | 0.74 |
| EC | 77.6 | 71.3 | 96.5 | 0.73 |
| Overall | ***97.0*** | ***90.0*** | ***96.8*** | ***0.76*** |
| **CELLO 2.5** | | | | |
| C | 97.9 | 95.7 | 95.1 | 0.86 |
| CM | 89.5 | 90.5 | 97.3 | 0.88 |
| CW | 100.0 | 31.8 | 99.2 | 0.56 |
| EC | 65.4 | 87.5 | 95.8 | 0.74 |
| Overall | ***93.7*** | ***93.7*** | ***96.9*** | ***0.76*** |
| **Gpos-PLoc*** | | | | |
| C | 91.5 | 99.1 | 92.1 | 0.76 |
| CM | 95.4 | 57.3 | 94.0 | 0.71 |
| CW | 66.7 | 18.2 | 99.0 | 0.34 |
| EC | 82.2 | 71.3 | 96.9 | 0.75 |
| Overall | ***91.2*** | ***90.7*** | ***95.5*** | ***0.64*** |
| **PA 2.5**** | | | | |
| C | 96.8 | 93.1 | 92.1 | 0.77 |
| CM | 93.5 | 11.5 | 88.0 | 0.30 |
| CW | N/A | N/A | N/A | N/A |
| EC | 49.0 | 87.5 | 92.5 | 0.62 |
| Overall | ***90.0*** | ***81.8*** | ***90.9*** | ***0.57*** |

†Abbreviations for localizations: C=cytoplasmic, CM=cytoplasmic membrane, CW=cell wall, EC=extracellular
†† Precision = TP / (TP + FP); Recall = TP / (TP + FN); Accuracy = (TP + TN) / (TP + FP + TN + FN);

$$MCC = \frac{TP \bullet TN - FP \bullet FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP = # true positives, FP = # false positives, TN = # true negatives, FN = # false negatives, MCC = Matthew's Coefficient Constant * *Software also predicts periplasmic SCL. None of the testing dataset proteins received a periplasmic SCL prediction.
** Software only predicts C, CM, EC categories. All proteins (including cell wall proteins) submitted to the server will receive one or more of these 3 localization predictions (or 'No predictions').

**Table 2: Performance comparison form Gram-negative bacterial SCL prediction software**

| Software† | Precision†† | Recall†† | Accuracy†† | MCC†† |
|---|---|---|---|---|
| **PSORTb 3.0** | | | | |
| C | 98.1 | 98.4 | 97.5 | 0.94 |
| CM | 96.4 | 91.5 | 97.6 | 0.92 |
| P | 91.3 | 72.4 | 99.0 | 0.81 |
| OM | 94.3 | 75.3 | 99.2 | 0.84 |
| EC | 92.0 | 61.3 | 98.3 | 0.74 |
| Overall | ***97.3*** | ***94.1*** | ***98.3*** | ***0.85*** |
| **PA 2.5** | | | | |
| C | 98.5 | 97.1 | 96.9 | 0.93 |
| CM | 97.7 | 82.1 | 96.0 | 0.87 |
| P | 76.1 | 89.7 | 98.8 | 0.82 |
| OM | 95.2 | 90.9 | 99.6 | 0.93 |
| EC | 92.2 | 52.4 | 98.0 | 0.69 |
| Overall | ***97.3*** | ***92.0*** | ***97.9*** | ***0.85*** |
| **PSORTb 2.0** | | | | |
| C | 98.3 | 94.0 | 94.7 | 0.88 |
| CM | 96.7 | 72.8 | 94.1 | 0.81 |
| P | 54.0 | 58.6 | 97.3 | 0.55 |
| OM | 79.1 | 59.1 | 98.5 | 0.68 |
| EC | 94.6 | 31.1 | 97.3 | 0.53 |
| Overall | ***95.9*** | ***85.3*** | ***96.3*** | ***0.69*** |
| **SubcellPredict*** | | | | |
| C | 97.3 | 98.5 | 95.9 | 0.41 |
| CM | N/A | N/A | N/A | N/A |
| P | 48.3 | 74.7 | 95.7 | 0.58 |
| OM | N/A | N/A | N/A | N/A |
| EC | 95.0 | 33.8 | 96.4 | 0.55 |
| Overall | ***94.3*** | ***94.3*** | ***96.0*** | ***0.52*** |
| **SLP-Local*** | | | | |
| C | 96.6 | 98.2 | 95.3 | 0.69 |
| CM | N/A | N/A | N/A | N/A |
| P | 52.5 | 78.7 | 96.4 | 0.63 |
| OM | N/A | N/A | N/A | N/A |
| EC | 81.9 | 26.2 | 96.0 | 0.45 |
| Overall | ***93.8*** | ***93.8*** | ***95.9*** | ***0.59*** |
| **Gneg-PLoc**** | | | | |
| C | 90.4 | 98.9 | 91.8 | 0.80 |
| CM | 95.0 | 74.4 | 94.1 | 0.81 |
| P | 56.5 | 50.0 | 97.3 | 0.52 |
| OM | 66.4 | 51.3 | 98.0 | 0.57 |
| EC | 80.0 | 39.1 | 97.3 | 0.55 |
| Overall | ***89.6*** | ***88.9*** | ***95.7*** | ***0.65*** |
| **CELLO 2.5** | | | | |
| C | 93.8 | 96.0 | 92.7 | 0.82 |
| CM | 95.9 | 72.5 | 93.9 | 0.80 |
| P | 35.2 | 81.0 | 94.9 | 0.51 |
| OM | 34.6 | 42.9 | 96.3 | 0.37 |
| EC | 69.2 | 44.9 | 97.1 | 0.54 |
| Overall | ***87.5*** | ***87.5*** | ***95.0*** | ***0.61*** |

†Abbreviations for localizations: C=cytoplasmic, CM=cytoplasmic membrane, P=periplasmic, OM=outer membrane, EC=extracellular

†† See Table 1 for definitions of precision, recall, accuracy and MCC.

*Software only predicts C, P, EC categories. All proteins (including membrane proteins) submitted to the server will receive one of these 3 SCL predictions

** Software also predicts flagellar, fimbrium and nucleoid localizations; however, none of test dataset proteins received one of these 3 SCL predictions

**Table 3: Performance comparison for archaeal proteins between PSORTb 3.0 archaeal option and software with Gram-positive SCL prediction capability**

| Software[†] | Precision[††] | Recall[††] | Accuracy[††] | MCC[††] |
|---|---|---|---|---|
| **PSORTb 3.0** | | | | |
| C | 98.1 | 96.9 | 95.8 | 0.85 |
| CM | 88.2 | 79.8 | 96.6 | 0.82 |
| CW | 100.0 | 83.3 | 99.6 | 0.91 |
| EC | 100.0 | 54.2 | 98.6 | 0.73 |
| Overall | ***97.2*** | ***93.4*** | ***97.7*** | ***0.83*** |
| **PSORTb 2.0** | | | | |
| C | 96.5 | 85.4 | 85.2 | 0.59 |
| CM | 93.5 | 69.0 | 96.1 | 0.78 |
| CW | 100.0 | 11.1 | 97.9 | 0.33 |
| EC | 77.8 | 58.3 | 98.2 | 0.66 |
| Overall | ***95.7*** | ***81.0*** | ***94.3*** | ***0.59*** |
| **Gpos-PLoc*** | | | | |
| C | 94.2 | 99.1 | 94.1 | 0.77 |
| CM | 91.4 | 63.1 | 95.3 | 0.74 |
| CW | 100.0 | 33.3 | 98.4 | 0.57 |
| EC | 48.3 | 58.3 | 96.7 | 0.51 |
| Overall | ***92.3*** | ***92.3*** | ***96.2*** | ***0.65*** |
| **PA 2.5**[**] | | | | |
| C | 98.4 | 87.9 | 88.4 | 0.66 |
| CM | 100.0 | 2.4 | 89.1 | 0.15 |
| CW | N/A | N/A | N/A | N/A |
| EC | 21.1 | 62.5 | 91.3 | 0.33 |
| Overall | ***90.0*** | ***77.5*** | ***89.6*** | ***0.38*** |
| **CELLO 2.5** | | | | |
| C | 95.6 | 91.9 | 89.7 | 0.66 |
| CM | 68.5 | 72.6 | 93.4 | 0.67 |
| CW | 100.0 | 5.6 | 97.8 | 0.23 |
| EC | 19.7 | 50.0 | 92.1 | 0.28 |
| Overall | ***86.5*** | ***86.5*** | ***93.2*** | ***0.46*** |

[†] Abbreviations for localizations are same as Table 1.

[††] See Table 1 for definitions of precision, recall, accuracy and MCC.

* Software also predicts periplasmic SCL. None of the testing dataset proteins received a periplasmic SCL prediction.

** Software only predicts C, CM, EC categories. All proteins (including cell wall proteins) submitted to the server will receive one or more of these 3 localization predictions (or ‘No prediction’).

**Table 4: List of proteins identified only in the cytoplasmic fraction and not periplasmic or extracellular fractions of Pseudomonas aeruginosa PA01. SCL predictions for PSORTb 2.0, PSORTb 2.0, PA 3.0, and PA 2.5 are shown below**

| Protein Description | PSORTb 3.0 Prediction | PA 3.0 Prediction | PA 2.5 predictions | PSORTb 2.0 Prediction |
|---|---|---|---|---|
| PA0296 probable glutamine synthetase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA0331 threonine dehydratase, biosynthetic | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA0363 phosphopantetheine adenylyltransferase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA0408 twitching motility protein PilG | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA0546 methionine adenosyltransferase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA0649 anthranilate synthase component II | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA0656 probable HIT family protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA0773 pyridoxal phosphate biosynthetic protein PdxJ | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA0903 alanyl-tRNA synthetase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA0916 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA1529 DNA ligase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA1756 3'-phosphoadenosine-5'-phosphosulfate reductase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA1769 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA1999 probable CoA transferase, subunit A | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA2586 response regulator GacA | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA2609 hypothetical protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA2666 probable 6-pyruvoyl tetrahydrobiopterin synthase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA2743 translation initiation factor IF-3 | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA3013 fatty-acid oxidation complex beta-subunit | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA3138 excinuclease ABC subunit B | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |

| | | | | |
|---|---|---|---|---|
| PA3170 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA3171 3-demethylubiquinone-9 3-methyltransferase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA3295 probable HIT family protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA3574 probable transcriptional regulator | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA3743 tRNA (guanine-N1)-methyltransferase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA3823 queuine tRNA-ribosyltransferase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA3940 probable DNA binding protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA4031 inorganic pyrophosphatase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA4247 50S ribosomal protein L18 | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA4249 30S ribosomal protein S8 | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA4443 ATP sulfurylase small subunit | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA4465 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA4626 glycerate dehydrogenase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA4645 probable purine/pyrimidine phosphoribosyl transferase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA4809 FdhE protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA4956 thiosulfate:cyanide sulfurtransferase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA4973 thiamin biosynthesis protein ThiC | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA5166 probable two-component response regulator | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA5323 acetylglutamate kinase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA5436 probable biotin carboxylase subunit of a transcarboxylase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA5453 GDP-mannose 4,6-dehydratase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA5459 hypothetical protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |

|  |  |  |  |  |
|---|---|---|---|---|
| PA5546 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic |
| PA4176 peptidyl-prolyl cis-trans isomerase C2 | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic,OuterMembrane |
| PA0761 L-aspartate oxidase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Cytoplasmic,Periplasmic |
| PA1135 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | CytoplasmicMembrane |
| PA4053 6,7-dimethyl-8-ribityllumazine synthase | Cytoplasmic | Cytoplasmic | Cytoplasmic | CytoplasmicMembrane |
| PA0036 tryptophan synthase beta chain | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA0066 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA0394 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA0932 cysteine synthase B | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA0937 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA1216 hypothetical protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA1576 probable 3-hydroxyisobutyrate dehydrogenase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA1750 phospho-2-dehydro-3-deoxyheptonate aldolase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA1757 homoserine kinase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA2111 hypothetical protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA2153 1,4-alpha-glucan branching enzyme | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA2618 hypothetical protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA3120 3-isopropylmalate dehydratase small subunit | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA3156 probable acetyltransferase WbpD | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA3263 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA3580 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA3753 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA4241 30S ribosomal protein S13 | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |

| | | | | |
|---|---|---|---|---|
| PA4246 30S ribosomal protein S5 | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA4258 50S ribosomal protein L22 | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA4259 30S ribosomal protein S19 | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA4261 50S ribosomal protein L23 | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA4433 50S ribosomal protein L13 | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA4563 30S ribosomal protein S20 | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA4567 50S ribosomal protein L27 | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA4696 acetolactate synthase large subunit | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA4715 probable aminotransferase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA4741 30S ribosomal protein S15 | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA4778 probable transcriptional regulator | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA4868 urease alpha subunit | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA4944 Hfq | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA5098 histidine ammonia-lyase | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA5334 ribonuclease PH | Cytoplasmic | Cytoplasmic | Cytoplasmic | Unknown |
| PA1294 ribonuclease D | Cytoplasmic | Cytoplasmic | Cytoplasmic, Extracellular | Cytoplasmic |
| PA3092 2,4-dienoyl-CoA reductase FadH1 | Cytoplasmic | Cytoplasmic | Cytoplasmic, Periplasmic | Cytoplasmic |
| PA2570 LecA | Cytoplasmic | Cytoplasmic | Periplasmic | Unknown |
| PA2717 chloroperoxidase precursor | Cytoplasmic | Cytoplasmic | Periplasmic, Extracellular | Unknown |
| PA0437 cytosine deaminase | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA0654 S-adenosylmethionine decarboxylase proenzyme | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA0715 hypothetical protein | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA0747 probable aldehyde dehydrogenase | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA0796 carboxyphosphonoenolpyruvate phosphonomutase | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |

|  |  |  |  |  |
|---|---|---|---|---|
| PA0832 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA0950 probable arsenate reductase | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA1124 deoxyguanosinetriphosphate triphosphohydrolase | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA1828 probable short-chain dehydrogenase | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA1900 probable phenazine biosynthesis protein | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA2443 L-serine dehydratase | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA2605 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA2630 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA2667 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA3043 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA3230 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA3255 hypothetical protein | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA3366 aliphatic amidase | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA3803 probable isoprenoid biosynthetic protein GcpE | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA4019 probable aromatic acid decarboxylase | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA4054 GTP cyclohydrolase II / 3,4-dihydroxy-2-butanone 4-phosphate synthase | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA4315 transcriptional regulator MvaT, P16 subunit | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA4445 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA4774 hypothetical protein | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA4880 probable bacterioferritin | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA5186 probable iron-containing alcohol dehydrogenase | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |
| PA5507 hypothetical protein | Cytoplasmic | Cytoplasmic | Unknown | Cytoplasmic |

| | | | | |
|---|---|---|---|---|
| PA5020 probable acyl-CoA dehydrogenase | Cytoplasmic | Cytoplasmic | Unknown | CytoplasmicMembrane |
| PA5060 polyhydroxyalkanoate synthesis protein PhaF | Cytoplasmic | Cytoplasmic | Unknown | CytoplasmicMembrane |
| PA0059 osmotically inducible protein OsmC | Cytoplasmic | Cytoplasmic | Unknown | Unknown |
| PA0083 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Unknown | Unknown |
| PA0449 hypothetical protein | Cytoplasmic | Cytoplasmic | Unknown | Unknown |
| PA0579 30S ribosomal protein S21 | Cytoplasmic | Cytoplasmic | Unknown | Unknown |
| PA1066 probable short-chain dehydrogenase | Cytoplasmic | Cytoplasmic | Unknown | Unknown |
| PA1204 conserved hypothetical protein | Cytoplasmic | Cytoplasmic | Unknown | Unknown |
| PA1358 hypothetical protein | Cytoplasmic | Cytoplasmic | Unknown | Unknown |
| PA2052 cyanate lyase | Cytoplasmic | Cytoplasmic | Unknown | Unknown |
| PA2764 hypothetical protein | Cytoplasmic | Cytoplasmic | Unknown | Unknown |
| PA2850 organic hydroperoxide resistance protein | Cytoplasmic | Cytoplasmic | Unknown | Unknown |
| PA3569 3-hydroxyisobutyrate dehydrogenase | Cytoplasmic | Cytoplasmic | Unknown | Unknown |
| PA3813 probable iron-binding protein IscU | Cytoplasmic | Cytoplasmic | Unknown | Unknown |
| PA3942 acyl-CoA thioesterase II | Cytoplasmic | Cytoplasmic | Unknown | Unknown |
| PA4250 30S ribosomal protein S14 | Cytoplasmic | Cytoplasmic | Unknown | Unknown |
| PA4255 50S ribosomal protein L29 | Cytoplasmic | Cytoplasmic | Unknown | Unknown |
| PA4640 malate:quinone oxidoreductase | Cytoplasmic | Cytoplasmic | Unknown | Unknown |
| PA4403 secretion protein SecA | Cytoplasmic | CytoplasmicMembrane | CytoplasmicMembrane | Cytoplasmic |
| PA4421 conserved hypothetical protein | Cytoplasmic | CytoplasmicMembrane | Unknown | Unknown |
| PA1642 selenophosphate synthetase | Cytoplasmic | Unknown | Cytoplasmic | Cytoplasmic |
| PA2003 3-hydroxybutyrate dehydrogenase | Cytoplasmic | Unknown | Cytoplasmic | Cytoplasmic |
| PA5313 probable pyridoxal-dependent aminotransferase | Cytoplasmic | Unknown | Cytoplasmic | Cytoplasmic |

| | | | | |
|---|---|---|---|---|
| PA5567 conserved hypothetical protein | Cytoplasmic | Unknown | Cytoplasmic | Cytoplasmic |
| PA1881 probable oxidoreductase | Cytoplasmic | Unknown | Periplasmic | Cytoplasmic |
| PA2379 probable oxidoreductase | Cytoplasmic | Unknown | Periplasmic | Cytoplasmic |
| PA0492 conserved hypothetical protein | Cytoplasmic | Unknown | Unknown | Cytoplasmic |
| PA0758 hypothetical protein | Cytoplasmic | Unknown | Unknown | Cytoplasmic |
| PA0900 hypothetical protein | Cytoplasmic | Unknown | Unknown | Cytoplasmic |
| PA2021 hypothetical protein | Cytoplasmic | Unknown | Unknown | Cytoplasmic |
| PA2557 probable AMP-binding enzyme | Cytoplasmic | Unknown | Unknown | Cytoplasmic |
| PA4010 hypothetical protein | Cytoplasmic | Unknown | Unknown | Cytoplasmic |
| PA4055 riboflavin synthase alpha chain | Cytoplasmic | Unknown | Unknown | Cytoplasmic |
| PA4115 conserved hypothetical protein | Cytoplasmic | Unknown | Unknown | Cytoplasmic |
| PA4558 probable peptidyl-prolyl cis-trans isomerase, FkbP-type | Cytoplasmic | Unknown | Unknown | Cytoplasmic |
| PA4918 hypothetical protein | Cytoplasmic | Unknown | Unknown | Cytoplasmic |
| PA5014 glutamate-ammonia-ligase adenylyltransferase | Cytoplasmic | Unknown | Unknown | Cytoplasmic |
| PA0004 DNA gyrase subunit B | Cytoplasmic | Unknown | Unknown | Unknown |
| PA0665 conserved hypothetical protein | Cytoplasmic | Unknown | Unknown | Unknown |
| PA1140 conserved hypothetical protein | Cytoplasmic | Unknown | Unknown | Unknown |
| PA1748 probable enoyl-CoA hydratase/isomerase | Cytoplasmic | Unknown | Unknown | Unknown |
| PA2504 hypothetical protein | Cytoplasmic | Unknown | Unknown | Unknown |
| PA3488 hypothetical protein | Cytoplasmic | Unknown | Unknown | Unknown |
| PA3539 conserved hypothetical protein | Cytoplasmic | Unknown | Unknown | Unknown |
| PA1430 transcriptional regulator LasR | Unknown | Cytoplasmic | Cytoplasmic | Unknown |
| PA1202 probable hydrolase | Unknown | Cytoplasmic | Unknown | Unknown |
| PA1768 hypothetical protein | Unknown | Cytoplasmic | Unknown | Unknown |

| | | | | |
|---|---|---|---|---|
| PA2222 hypothetical protein | Unknown | Cytoplasmic | Unknown | Unknown |
| PA2580 conserved hypothetical protein | Unknown | Cytoplasmic | Unknown | Unknown |
| PA2871 hypothetical protein | Unknown | Cytoplasmic | Unknown | Unknown |
| PA3421 conserved hypothetical protein | Unknown | OuterMembrane | Unknown | Unknown |
| PA0974 conserved hypothetical protein | Unknown | Unknown | Unknown | Unknown |
| PA4588 glutamate dehydrogenase | Cytoplasmic, OuterMembrane | Cytoplasmic | Unknown | Cytoplasmic |
| PA2826 probable glutathione peroxidase | Cytoplasmic, Periplasmic | Cytoplasmic | Unknown | Periplasmic |
| PA3481 conserved hypothetical protein | CytoplasmicMembrane | Cytoplasmic | Cytoplasmic | Unknown |
| PA0143 purine nucleosidase Nuh | Non-Cytoplasmic | Cytoplasmic | Unknown | Non-Cytoplasmic |
| PA2812 probable ATP-binding component of ABC transporter | CytoplasmicMembrane | CytoplasmicMembrane | CytoplasmicMembrane | CytoplasmicMembrane |
| PA3839 probable sodium:sulfate symporter | CytoplasmicMembrane | CytoplasmicMembrane | CytoplasmicMembrane | CytoplasmicMembrane |
| PA1608 probable chemotaxis transducer | CytoplasmicMembrane, OuterMembrane | OuterMembrane | CytoplasmicMembrane, Extracellular | CytoplasmicMembrane |
| PA2300 chitinase | Extracellular | Unknown | Periplasmic, Extracellular | Extracellular |

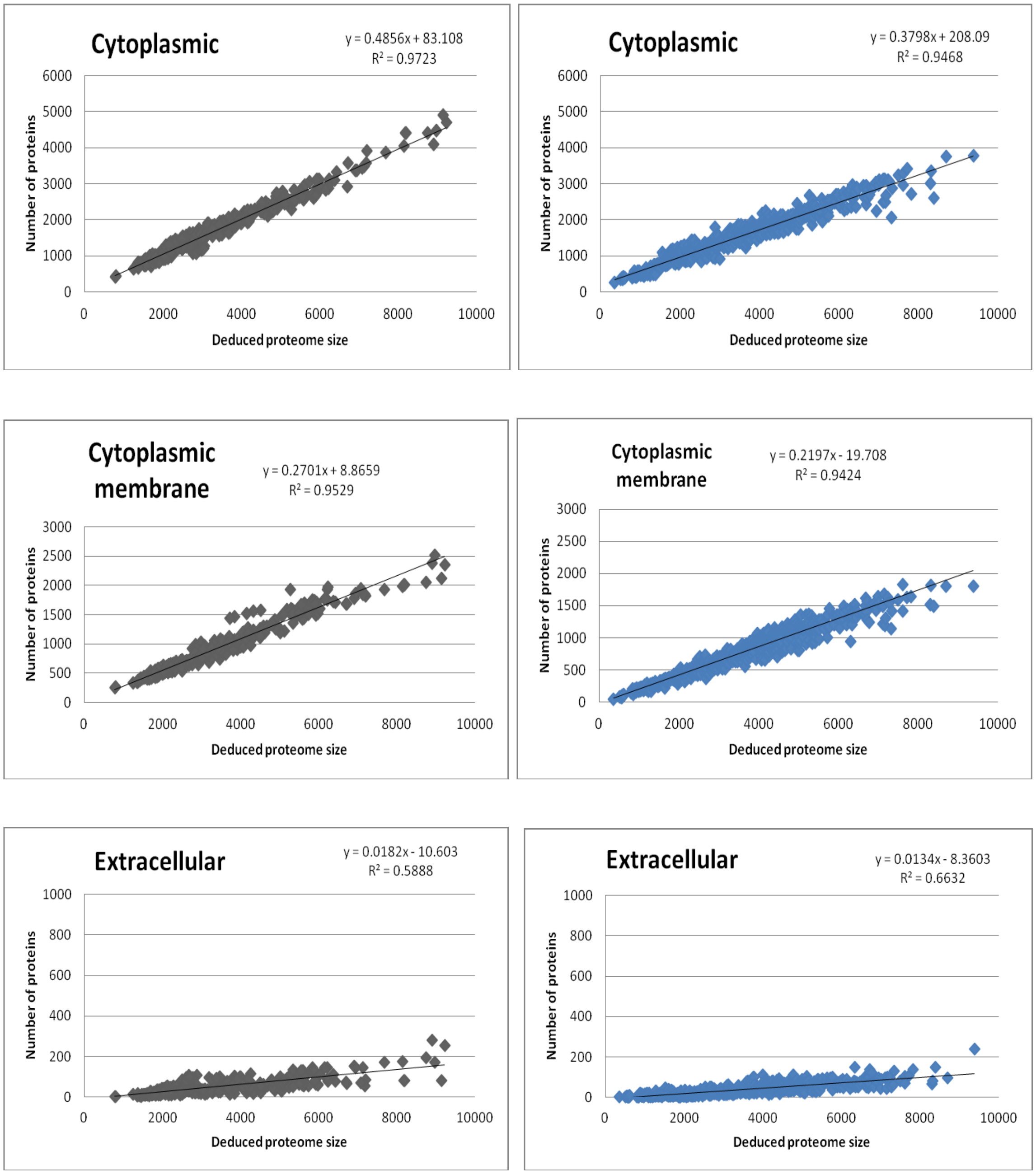
# Appendix B - Supplementary data for chapter 3

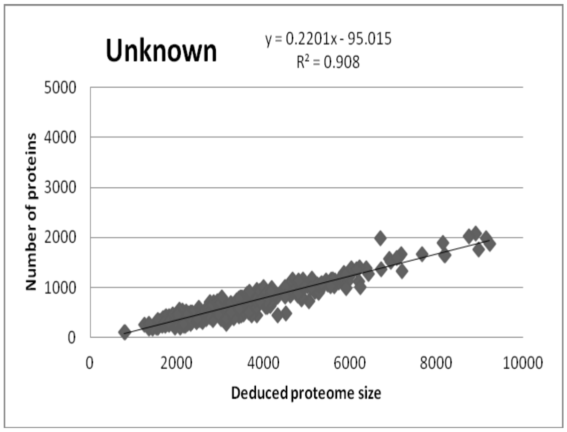
**Raw data for outer membrane prediction using phylum and BLAST results of OMP85 gene from *Neisseri gonorrhoeae*, *Synechococcus sp*., *Thermosipho africanus*, and *Thermus thermophilus* against bacterial genomes – please see Excel file - OMP85_predictions.xlsx**

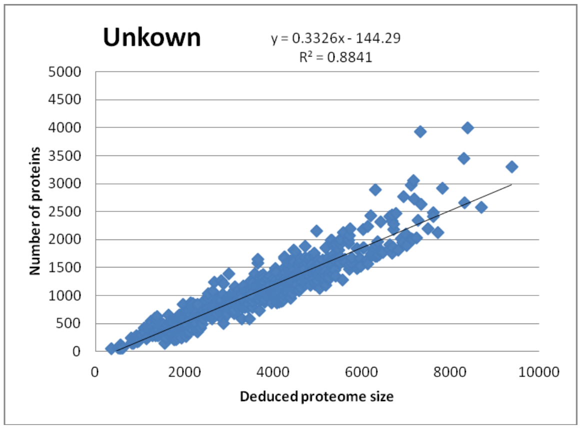
# Appendix C - Supplementary data for chapter 4

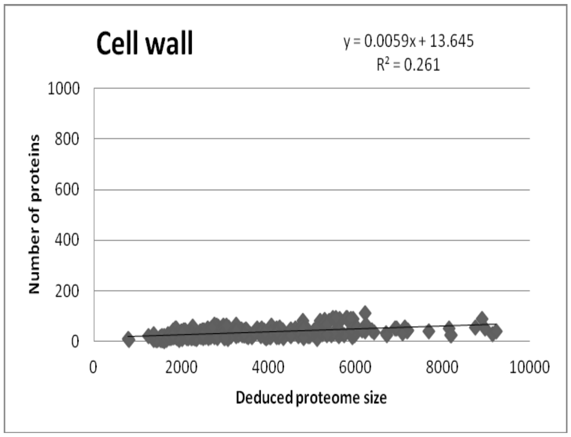
**Figure 1: Linear trends for Number of proteins vs. proteome size for all SCLs for a) monoderms b) diderms**

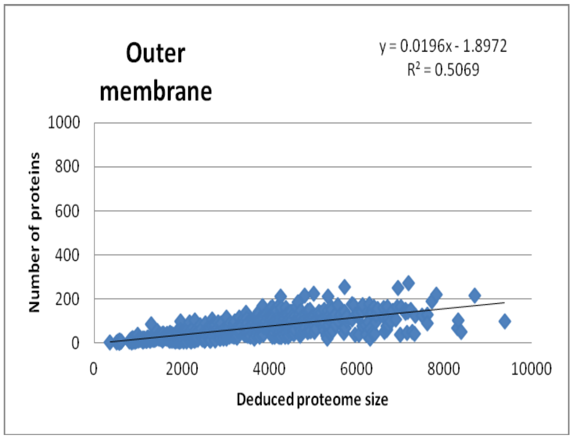
a) b)




Unknown
y = 0.2201x - 95.015
R² = 0.908
Number of proteins
5000
4000
3000
2000
1000
0
0
2000
4000
6000
8000
10000
Deduced proteome size


Unkown
y = 0.3326x - 144.29
R² = 0.8841
Number of proteins
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
2000
4000
6000
8000
10000
Deduced proteome size


Cell wall
y = 0.0059x + 13.645
R² = 0.261
Number of proteins
1000
800
600
400
200
0
0
2000
4000
6000
8000
10000
Deduced proteome size


Outer
membrane
y = 0.0196x - 1.8972
R² = 0.5069
Number of proteins
1000
800
600
400
200
0
0
2000
4000
6000
8000
10000
Deduced proteome size

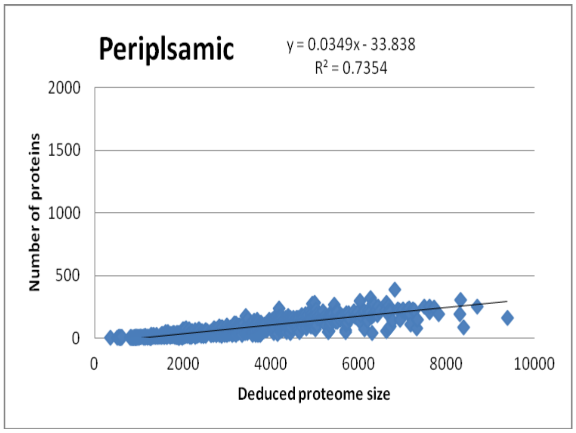

Periplsamic
y = 0.0349x - 33.838
R² = 0.7354
Number of proteins
2000
1500
1000
500
0
0
2000
4000
6000
8000
10000
Deduced proteome size

**Figure 2. Linear trends for number of proteins vs. proteome size for Archaea**

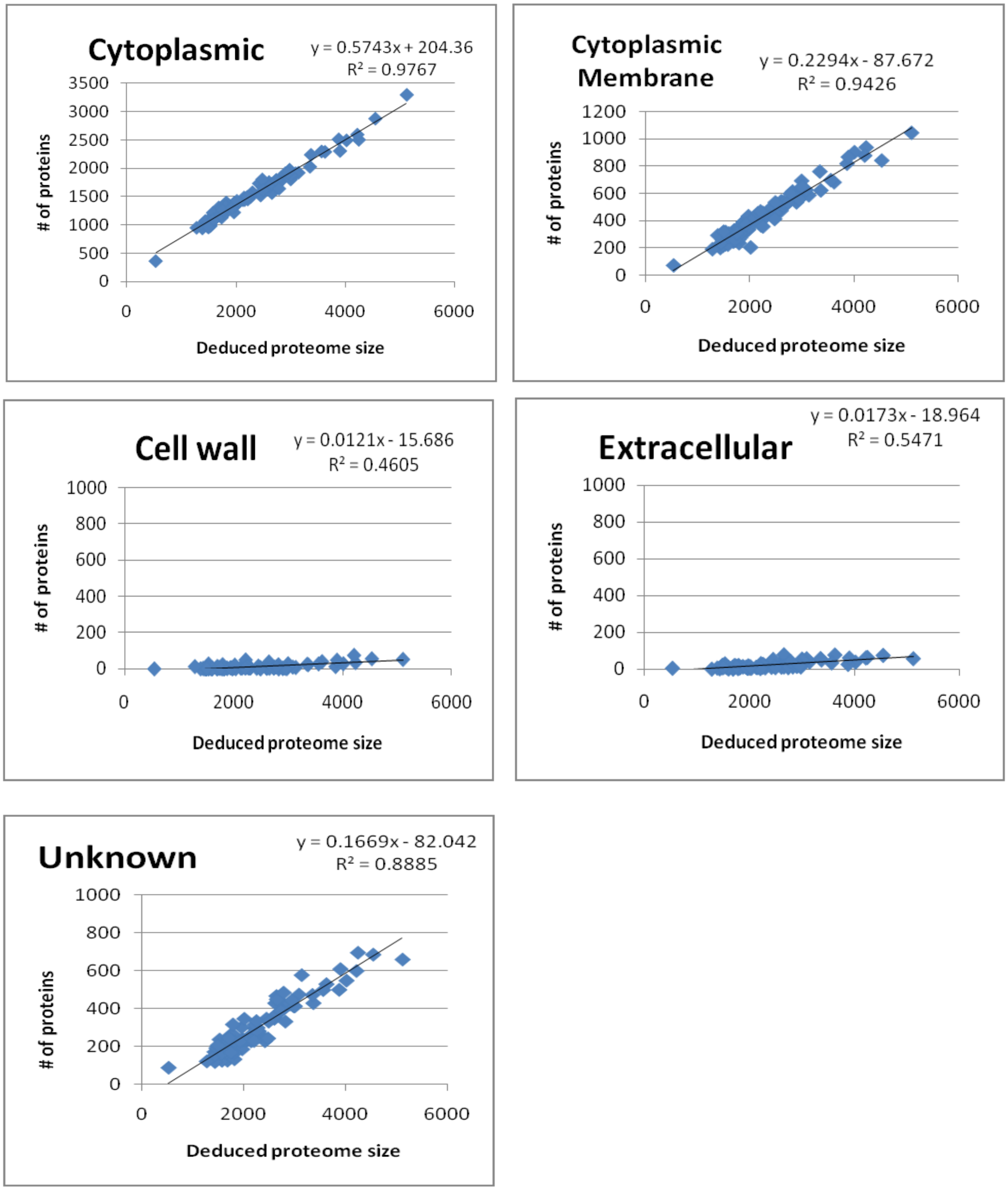


**Table 1. Number of predictions for each SCL for monodermic bacterial proteomes – please see Excel file SCLcounts.xlsx**

**Table 2. Number of predictions for each SCL for didermic bacterial proteomes – please see Excel file SCLcounts.xlsx**

**Table 3. Number of predictions for each SCL for archaeal proteomes**

| GenomeName | C | CM | CW | EC | Unknown | Total |
|---|---|---|---|---|---|---|
| Aeropyrum_pernix | 1171 | 331 | 3 | 6 | 189 | 1700 |
| Methanobacterium_thermoautotrophicum | 1349 | 325 | 6 | 14 | 179 | 1873 |
| Archaeoglobus_fulgidus | 1735 | 443 | 4 | 10 | 228 | 2420 |
| Pyrococcus_horikoshii | 1323 | 411 | 8 | 7 | 206 | 1955 |
| Methanococcus_jannaschii | 1333 | 265 | 2 | 9 | 177 | 1786 |
| Pyrococcus_abyssi | 1257 | 353 | 5 | 3 | 164 | 1782 |
| Thermoplasma_acidophilum | 976 | 314 | 1 | 18 | 173 | 1482 |
| Halobacterium_sp | 1627 | 513 | 11 | 41 | 430 | 2622 |
| Thermoplasma_volcanium | 984 | 298 | 1 | 15 | 201 | 1499 |
| Sulfolobus_solfataricus | 1976 | 572 | 3 | 12 | 414 | 2977 |
| Sulfolobus_tokodaii | 1783 | 617 | 1 | 12 | 412 | 2825 |
| Pyrobaculum_aerophilum | 1759 | 476 | 5 | 16 | 347 | 2603 |
| Pyrococcus_furiosus | 1440 | 412 | 6 | 10 | 257 | 2125 |
| Methanopyrus_kandleri | 1311 | 249 | 2 | 0 | 125 | 1687 |
| Methanosarcina_acetivorans | 2875 | 844 | 58 | 75 | 688 | 4540 |
| Methanosarcina_mazei | 2238 | 625 | 25 | 50 | 430 | 3368 |
| Nanoarchaeum_equitans | 368 | 75 | 1 | 5 | 87 | 536 |
| Methanococcus_maripaludis_S2 | 1206 | 298 | 6 | 11 | 201 | 1722 |
| Picrophilus_torridus_DSM_9790 | 985 | 320 | 1 | 23 | 206 | 1535 |
| Haloarcula_marismortui_ATCC_43049 | 2502 | 940 | 36 | 64 | 698 | 4240 |
| Thermococcus_kodakaraensis_KOD1 | 1557 | 459 | 7 | 15 | 268 | 2306 |
| Sulfolobus_acidocaldarius_DSM_639 | 1482 | 425 | 9 | 11 | 296 | 2223 |
| Methanosarcina_barkeri_fusaro | 2290 | 684 | 42 | 77 | 531 | 3624 |
| Natronomonas_pharaonis | 1835 | 606 | 24 | 23 | 332 | 2820 |
| Methanosphaera_stadtmanae | 1009 | 230 | 29 | 29 | 237 | 1534 |
| Methanospirillum_hungatei_JF-1 | 1923 | 589 | 9 | 39 | 579 | 3139 |
| Methanococcoides_burtonii_DSM_6242 | 1503 | 443 | 6 | 22 | 299 | 2273 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Haloquadratum_walsbyi | 1609 | 511 | 40 | 18 | 468 | 2646 |
| Methanosaeta_thermophila_PT | 1210 | 263 | 15 | 13 | 195 | 1696 |
| Thermofilum_pendens_Hrk_5 | 1291 | 393 | 3 | 12 | 177 | 1876 |
| Pyrobaculum_islandicum_DSM_4184 | 1383 | 341 | 8 | 6 | 240 | 1978 |
| Hyperthermus_butylicus | 1177 | 260 | 1 | 2 | 162 | 1602 |
| Methanocorpusculum_labreanum_Z | 1134 | 320 | 8 | 17 | 260 | 1739 |
| Staphylothermus_marinus_F1 | 1077 | 303 | 5 | 9 | 176 | 1570 |
| Methanoculleus_marisnigri_JR1 | 1667 | 457 | 9 | 20 | 336 | 2489 |
| Pyrobaculum_calidifontis_JCM_11548 | 1487 | 421 | 5 | 8 | 228 | 2149 |
| Methanococcus_maripaludis_C5 | 1282 | 304 | 8 | 9 | 219 | 1822 |
| Pyrobaculum_arsenaticum_DSM_13514 | 1578 | 427 | 6 | 7 | 281 | 2299 |
| Metallosphaera_sedula_DSM_5348 | 1493 | 460 | 9 | 12 | 282 | 2256 |
| uncultured_methanogenic_archaeon_RC-I | 1913 | 628 | 11 | 57 | 476 | 3085 |
| Methanobrevibacter_smithii_ATCC_35061 | 1202 | 282 | 25 | 20 | 264 | 1793 |
| Methanococcus_vannielii_SB | 1199 | 264 | 6 | 9 | 200 | 1678 |
| Methanococcus_aeolicus_Nankai-3 | 1083 | 219 | 9 | 22 | 157 | 1490 |
| Methanococcus_maripaludis_C7 | 1243 | 312 | 11 | 15 | 207 | 1788 |
| Candidatus_Methanoregula_boonei_6A8 | 1522 | 510 | 17 | 53 | 348 | 2450 |
| Ignicoccus_hospitalis_KIN4_I | 1055 | 202 | 4 | 2 | 171 | 1434 |
| Caldivirga_maquilingensis_IC-167 | 1224 | 423 | 5 | 10 | 301 | 1963 |
| Methanococcus_maripaludis_C6 | 1290 | 294 | 6 | 11 | 225 | 1826 |
| Nitrosopumilus_maritimus_SCM1 | 1211 | 238 | 15 | 14 | 317 | 1795 |
| Halobacterium_salinarum_R1 | 1713 | 545 | 11 | 46 | 434 | 2749 |
| Candidatus_Korarchaeum_cryptofilum_OPF8 | 1100 | 309 | 6 | 3 | 184 | 1602 |
| Thermoproteus_neutrophilus_V24Sta | 1391 | 334 | 4 | 5 | 232 | 1966 |
| Thermococcus_onnurineus_NA1 | 1334 | 436 | 2 | 18 | 185 | 1975 |
| Desulfurococcus_kamchatkensis_1221n | 1006 | 264 | 2 | 8 | 191 | 1471 |
| Candidatus_Methanosphaerula_palustris_E1_9c | 1565 | 535 | 27 | 78 | 450 | 2655 |
| Halorubrum_lacusprofundi_ATCC_49239 | 2297 | 700 | 29 | 34 | 500 | 3560 |
| Sulfolobus_islandicus_M_14_25 | 1689 | 545 | 5 | 11 | 358 | 2608 |
| Sulfolobus_islandicus_L_S_2_15 | 1798 | 535 | 7 | 8 | 389 | 2737 |
| Sulfolobus_islandicus_Y_G_57_14 | 1891 | 568 | 4 | 17 | 423 | 2903 |
| Sulfolobus_islandicus_Y_N_15_51 | 1914 | 537 | 4 | 15 | 430 | 2900 |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Sulfolobus_islandicus_M_16_27 | 1725 | 540 | 5 | 12 | 374 | 2656 |
| Sulfolobus_islandicus_M_16_4 | 1791 | 537 | 5 | 13 | 389 | 2735 |
| Thermococcus_gammatolerans_EJ3 | 1442 | 457 | 12 | 12 | 233 | 2156 |
| Thermococcus_sibiricus_MM_739 | 1391 | 404 | 2 | 5 | 233 | 2035 |
| Methanocaldococcus_fervens_AG86 | 1218 | 227 | 3 | 9 | 124 | 1581 |
| Halorhabdus_utahensis_DSM_12940 | 1804 | 695 | 30 | 55 | 414 | 2998 |
| Halomicrobium_mukohataei_DSM_12286 | 2027 | 763 | 27 | 58 | 474 | 3349 |
| Methanocaldococcus_vulcanius_M7 | 1301 | 274 | 4 | 16 | 147 | 1742 |
| Methanocella_paludicola_SANAE | 1877 | 652 | 14 | 48 | 413 | 3004 |
| Archaeoglobus_profundus_DSM_5631_uid43493 | 1400 | 279 | 1 | 11 | 132 | 1823 |
| Haloterrigena_turkmenica_DSM_5511_uid43501 | 3293 | 1047 | 54 | 57 | 662 | 5113 |
| Sulfolobus_islandicus_L_D_8_5_uid43679 | 1921 | 559 | 4 | 19 | 445 | 2948 |
| Methanobrevibacter_ruminantium_M1_uid45857 | 1453 | 368 | 51 | 30 | 315 | 2217 |
| Ferroglobus_placidus_DSM_10642_uid40863 | 1804 | 415 | 4 | 14 | 243 | 2480 |
| Methanocaldococcus_FS406_22_uid42499 | 1331 | 308 | 3 | 8 | 166 | 1816 |
| Natrialba_magadii_ATCC_43099_uid46245 | 2593 | 880 | 77 | 61 | 601 | 4212 |
| Aciduliprofundum_boonei_T469_uid43333 | 1071 | 295 | 10 | 13 | 155 | 1544 |
| Haloferax_volcanii_DS2_uid46845 | 2492 | 905 | 30 | 38 | 550 | 4015 |
| Methanohalophilus_mahii_DSM_5219_uid47313 | 1371 | 360 | 7 | 16 | 233 | 1987 |
| Methanocaldococcus_infernus_ME_uid48803 | 1074 | 238 | 3 | 8 | 118 | 1441 |
| Thermosphaera_aggregans_DSM_11486_uid48993 | 943 | 295 | 2 | 6 | 141 | 1387 |
| Staphylothermus_hellenicus_DSM_12710_uid45893 | 1098 | 283 | 2 | 10 | 206 | 1599 |
| Methanococcus_voltae_A3_uid49529 | 1185 | 259 | 9 | 20 | 244 | 1717 |
| Methanohalobium_evestigatum_Z_7303_uid49857 | 1519 | 360 | 14 | 26 | 335 | 2254 |
| Halalkalicoccus_jeotgali_B3_uid50305 | 2511 | 821 | 14 | 26 | 501 | 3873 |
| Acidilobus_saccharovorans_345_15_uid51395 | 960 | 322 | 3 | 8 | 206 | 1499 |
| Methanothermobacter_marburgensis_Marburg_uid51637 | 1261 | 310 | 5 | 13 | 168 | 1757 |
| Ignisphaera_aggregans_DSM_17230_uid51875 | 1308 | 378 | 7 | 11 | 226 | 1930 |
| Methanoplanus_petrolearius_DSM_11571_uid52695 | 1643 | 601 | 13 | 42 | 486 | 2785 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Vulcanisaeta_distributa_DSM_14429_uid52827 | 1614 | 537 | 2 | 8 | 332 | 2493 |
| Methanothermus_fervidus_DSM_2088_uid60167 | 954 | 194 | 14 | 0 | 121 | 1283 |
| Halogeometricum_borinquense_DSM_11551_uid54919 | 2306 | 870 | 50 | 61 | 611 | 3898 |
| Thermococcus_barophilus_MP_uid54733 | 1498 | 471 | 6 | 3 | 229 | 2207 |
| Cenarchaeum_symbiosum_A_uid61411 | 1428 | 207 | 22 | 13 | 347 | 2017 |

# Reference List

Abdallah, A.M., Gey van Pittius, N.C., Champion, P.A., Cox, J., Luirink, J., Vandenbroucke-Grauls, C.M., et al. (2007) Type VII secretion--mycobacteria show the way. *Nat.Rev.Microbiol.*, 5:883-891.

Akopyan, K., Edgren, T., Wang-Edgren, H., Rosqvist, R., Fahlgren, A., Wolf-Watz, H., et al. (2011) Translocation of surface-localized effectors in type III secretion. *Proc.Natl.Acad.Sci.U.S.A.*, 108:1639-1644.

Albers, S.V. and Meyer, B.H. (2011) The archaeal cell envelope. *Nat.Rev.Microbiol.*, 9:414-426.

Aldridge, P. and Hughes, K.T. (2002) Regulation of flagellar assembly. *Curr. Opin. Microbiol.*, 5:160-165.

Alvarez-Martinez, C.E. and Christie, P.J. (2009) Biological diversity of prokaryotic type IV secretion systems. *Microbiol.Mol.Biol.Rev.*, 73:775-808.

Anderson, D.M. and Schneewind, O. (1997) A mRNA signal for the type III secretion of Yop proteins by Yersinia enterocolitica. *Science*, 278:1140-1143.

Arnold, R., Brandmaier, S., Kleine, F., Tischler, P., Heinz, E., Behrens, S., et al. (2009) Sequence-based prediction of type III secreted proteins. *PLoS Pathog.*, 5:e1000376.

Bagos, P.G., Tsirigos, K.D., Plessas, S.K., Liakopoulos, T.D. and Hamodrakas, S.J. (2009) Prediction of signal peptides in archaea. *Protein Eng.Des.Sel.*, 22:27-35.

Beard, S.J., Handley, B.A., Hayes, P.K. and Walsby, A.E. (1999) The diversity of gas vesicle genes in Planktothrix rubescens from Lake Zurich. *Microbiology*, 145 ( Pt 10):2757-2768.

Bechtluft, P., Nouwen, N., Tans, S.J. and Driessen, A.J. (2010) SecB--a chaperone dedicated to protein translocation. *Mol.Biosyst*, 6:620-627.

Bendtsen, J.D., Jensen, L.J., Blom, N., von Heijne, G., and Brunak, S. (2004a) Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng. Des. Sel.*, 17:349-356.

Bendtsen, J.D., Nielsen, H., von Heijne, G. and Brunak, S. (2004b) Improved prediction of signal peptides: SignalP 3.0. *J.Mol.Biol.*, 340:783-795.

Berg, H.C. and Anderson, R.A. (1973) Bacteria swim by rotating their flagellar filaments. *Nature*, 245:380-382.

Berks, B.C., Palmer, T. and Sargent, F. (2003) The Tat protein translocation pathway and its role in microbial physiology. *Adv.Microb.Physiol.*, 47:187-254.

Beveridge, T.J. (1990) Mechanism of gram variability in select bacteria. *J. Bacteriol.* 172:1609-1620.

Beveridge, T.J. (2001) Use of the gram stain in microbiology. *Biotech.Histochem.*, 76:111-118.

Bhasin, M., Garg, A. and Raghava, G.P. (2005) PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics*, 21:2522-2524.

Billion, A., Ghai, R., Chakraborty, T. and Hain, T. (2006) Augur--a computational pipeline for whole genome microbial surface protein prediction and classification. *Bioinformatics*, 22:2819-2820.

Bingle, L.E., Bailey, C.M. and Pallen, M.J. (2008) Type VI secretion: a beginner's guide. *Curr.Opin.Microbiol.*, 11:3-8.

Bomberger, J.M., Maceachran, D.P., Coutermarsh, B.A., Ye, S., O'Toole, G.A., and Stanton, B.A. (2009) Long-distance delivery of bacterial virulence factors by Pseudomonas aeruginosa outer membrane vesicles. *PLoS Pathog.*, 5:e1000382.

Brown, D.R. and Parker, C.D. (1987) Cloning of the filamentous hemagglutinin of Bordetella pertussis and its expression in Escherichia coli. *Infect.Immun.*, 55:154-161.

Brutinel, E.D. and Yahr, T.L. (2008) Control of gene expression by type III secretion activity. *Curr. Opin. Microbiol.*, 11:128-133.

Bulashevska, A. and Eils, R. (2006) Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. *BMC Bioinformatics*, 7:298.

Burts, M.L., Williams, W.A., DeBord, K. and Missiakas, D.M. (2005) EsxA and EsxB are secreted by an ESAT-6-like system that is required for the pathogenesis of Staphylococcus aureus infections. *Proc.Natl.Acad.Sci.U.S.A.*, 102:1169-1174.

Cascales, E. (2008) The type VI secretion toolkit. *EMBO Rep.*, 9:735-741.

Cascales, E. and Christie, P.J. (2003) The versatile bacterial type IV secretion systems. *Nat.Rev.Microbiol.*, 1:137-149.

Chan, Q.W.T., Howes, C.G. and Foster, L.J. (2006) Quantitative comparison of caste differences in honeybee hemolymph. *Mol Cell Proteomics*, 5:2252-2262.

Chang, J.M., Su, E.C., Lo, A., Chiu, H.S., Sung, T.Y. and Hsu, W.L. (2008) PSLDoc: Protein subcellular localization prediction based on gapped-dipeptides and probabilistic latent semantic analysis. *Proteins*, 72:693-710.

Chou, K.C. (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, 21:10-19.

Chou, K.C. and Shen, H.B. (2006) Large-scale predictions of gram-negative bacterial protein subcellular locations. *J.Proteome Res.*, 5:3420-3428.

Christie, P.J. (2009) Structural biology: Translocation chamber's secrets. *Nature*, 462:992-994.

Christie, P.J. and Cascales, E. (2005) Structural and dynamic properties of bacterial type IV secretion systems (review). *Mol.Membr.Biol.*, 22:51-61.

Collazo, C.M. and Galan, J.E. (1996) Requirement for exported proteins in secretion through the invasion-associated type III system of Salmonella typhimurium. *Infect.Immun.*, 64:3524-3531.

Cornelis, G.R. (2006) The type III secretion injectisome. *Nat.Rev.Microbiol.*, 4:811-825.

Coulthurst, S.J. and Palmer, T. (2008) A new way out: protein localization on the bacterial cell surface via Tat and a novel Type II secretion system. *Mol.Microbiol.*, 69:1331-1335.

Cristobal, S., de Gier, J.W., Nielsen, H. and von Heijne, G. (1999) Competition between Sec- and TAT-dependent protein translocation in Escherichia coli. *EMBO J.*, 18:2982-2990.

Dean, J.E., Abrusci, P., Johnson, S., and Lea, S.M. (2009) Timing is everything: the regulation of type III secretion. *Cell. Mol. Life Sci.*, 67:1065-1075.

de Champdore, M., Staiano, M., Rossi, M. and D'Auria, S. (2007) Proteins from extremophiles as stable tools for advanced biotechnological applications of high social interest. *J.R.Soc.Interface*, 4:183-191.

de Jong, M.F., Sun, Y.H., den Hartigh, A.B., van Dijl, J.M. and Tsolis, R.M. (2008) Identification of VceA and VceC, two members of the VjbR regulon that are translocated into macrophages by the Brucella type IV secretion system. *Mol.Microbiol.*, 70:1378-1396.

Desvaux, M., Parham, N.J. and Henderson, I.R. (2004) Type V protein secretion: simplicity gone awry? *Curr.Issues Mol.Biol.*, 6:111-124.

Driessen, A.J. (2001) SecB, a molecular chaperone with two faces. *Trends Microbiol.*, 9:193-196.

Dubey, G.P. and Ben-Yehuda, S. (2011) Intercellular nanotubes mediate bacterial communication. *Cell*, 144:590-600.

Ellen, A.F., Albers, S.V., Huibers, W., Pitcher, A., Hobel, C.F., Schwarz, H., et al. (2009) Proteomic analysis of secreted membrane vesicles of archaeal Sulfolobus species reveals the presence of endosome sorting complex components. *Extremophiles*, 13:67-79.

Ellis, T.N. and Kuehn, M.J. (2010) Virulence and immunomodulatory roles of bacterial outer membrane vesicles. *Microbiol. Mol. Biol. Rev.*, 74:81-94.

Evans, E.L. and Allen, M.M. (1973) Phycobilisomes in Anacystis nidulans. *J. Bacteriol.*, 113:403-408.

Fagerlund, R.D. and Eaton-Rye, J.J. (2011) The lipoproteins of cyanobacterial photosystem II. *J.Photochem.Photobiol.B.*, 104:191-203.

Fekkes, P. and Driessen, A.J. (1999) Protein targeting to the bacterial cytoplasmic membrane. *Microbiol.Mol. Biol. Rev.,* 63:161-173.

Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F. and Whitehouse, C.M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246:64-71.

Filloux, A. (2004) The underlying mechanisms of type II protein secretion. *Biochim.Biophys.Acta*, 1694:163-179.

Franke, C.M., Tiemersma, J., Venema, G. and Kok, J. (1999) Membrane topology of the lactococcal bacteriocin ATP-binding cassette transporter protein LcnC. Involvement of LcnC in lactococcin a maturation. *J.Biol.Chem.*, 274:8484-8490.

Fredrickson, J.K., Zachara, J.M., Balkwill, D.L., Kennedy, D., Li, S.M., Kostandarithes, H.M., et al. (2004) Geomicrobiology of high-level nuclear waste-contaminated vadose sediments at the hanford site, washington state. *Appl.Environ.Microbiol.*, 70:4230-4241.

Frigaard, N.U. and Bryant, D.A. (2004) Seeing green bacteria in a new light: genomics-enabled studies of the photosynthetic apparatus in green sulfur bacteria and filamentous anoxygenic phototrophic bacteria. *Arch.Microbiol.*, 182:265-276.

Frols, S., Ajon, M., Wagner, M., Teichmann, D., Zolghadr, B., Folea, M., et al. (2008) UV-inducible cellular aggregation of the hyperthermophilic archaeon Sulfolobus solfataricus is mediated by pili formation. *Mol.Microbiol.*, 70:938-952.

Fuerst, J.A. and Sagulenko, E. (2011) Beyond the bacterium: planctomycetes challenge our concepts of microbial structure and function. *Nat.Rev.Microbiol.*, 9:403-413.

Fullner, K.J., Lara, J.C. and Nester, E.W. (1996) Pilus assembly by Agrobacterium T-DNA transfer genes. *Science*, 273:1107-1109.

Gardy, J.L. (2003) PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res.*, 31:3613-3617.

Gardy, J.L., Laird, M.R., Chen, F., Rey, S., Walsh, C.J., Ester, M., et al. (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*, 21:617-623.

Gardy, J.L. and Brinkman, F.S.L. (2006) Methods for predicting bacterial protein subcellular localization. *Nat Rev Micro*, 4:741-751.

Garufi, G., Butler, E. and Missiakas, D. (2008) ESAT-6-like protein secretion in Bacillus anthracis. *J.Bacteriol.*, 190:7004-7011.

Ghosh, P. (2004) Process of protein transport by the type III secretion system. *Microbiol. Mol. Biol. Rev.*, 68:771-795.

Gimenez, M.I., Dilks, K. and Pohlschroder, M. (2007) Haloferax volcanii twin-arginine translocation substates include secreted soluble, C-terminally anchored and lipoproteins. *Mol.Microbiol.*, 66:1597-1606.

Goudenege, D., Avner, S., Lucchetti-Miganeh, C. and Barloy-Hubler, F. (2010) CoBaltDB: Complete bacterial and archaeal orfeomes subcellular localization database and associated resources. *BMC Microbiol.*, 10:88.

Graham, J.E., Clark, M.E., Nadler, D.C., Huffer, S., Chokhawala, H.A., Rowland, S.E., et al. (2011) Identification and characterization of a multidomain hyperthermophilic cellulase from an archaeal enrichment. *Nat.Commun.*, 2:375.

Griffiths, W.J., Jonsson, A.P., Liu, S., Rai, D.K. and Wang, Y. (2001) Electrospray and tandem mass spectrometry in biochemistry. *Biochem.J.*, 355:545-561.

Grosskinsky, U., Schutz, M., Fritz, M., Schmid, Y., Lamparter, M.C., Szczesny, P., et al. (2007) A conserved glycine residue of trimeric autotransporter domains plays a key role in Yersinia adhesin A autotransport. *J.Bacteriol.*, 189:9011-9019.

Guo, J., Lin, Y. and Liu, X. (2006) GNBSL: a new integrative system to predict the subcellular location for Gram-negative bacteria proteins. *Proteomics*, 6:5099-5105.

Guo, T., Hua, S., Ji, X. and Sun, Z. (2004) DBSubLoc: database of protein subcellular localization. *Nucleic Acids Res.*, 32:D122-4.

Henderson, I.R., Navarro-Garcia, F. and Nataro, J.P. (1998) The great escape: structure and function of the autotransporter proteins. *Trends Microbiol.*, 6:370-378.

Heveker, N., Bonnaffe, D. and Ullmann, A. (1994) Chemical fatty acylation confers hemolytic and toxic activities to adenylate cyclase protixin of Bordetella Pertussis. *J. Biol. Chem.*, 269:32844-32847.

Holland, I.B., Schmitt, L. and Young, J. (2005) Type 1 protein secretion in bacteria, the ABC-transporter dependent pathway (review). *Mol.Membr.Biol.*, 22:29-39.

Horn, C., Paulmann, B., Kerlen, G., Junker, N. and Huber, H. (1999) In vivo observation of cell division of anaerobic hyperthermophiles by using a high-intensity dark-field microscope. *J.Bacteriol.*, 181:5114-5118.

Hu, Q., Noll, R.J., Li, H., Makarov, A., Hardman, M. and Graham Cooks, R. (2005) The Orbitrap: a new mass spectrometer. *J.Mass Spectrom.*, 40:430-443.

Hua, S. and Sun, Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17:721-728.

Huber, H., Burggraf, S., Mayer, T., Wyschkony, I., Rachel, R. and Stetter, K.O. (2000) Ignicoccus gen. nov., a novel genus of hyperthermophilic, chemolithoautotrophic Archaea, represented by two new species, Ignicoccus islandicus sp nov and Ignicoccus pacificus sp nov. and Ignicoccus pacificus sp. nov. *Int.J.Syst.Evol.Microbiol.*, 50 Pt 6:2093-2100.

Imai, K., Asakawa, N., Tsuji, T., Akazawa, F., Ino, A., Sonoyama, M., et al. (2008) SOSUI-GramN: high performance prediction for sub-cellular localization of proteins in gram-negative bacteria. *Bioinformation*, 2:417-421.

Johnson, T.L., Abendroth, J., Hol, W.G. and Sandkvist, M. (2006) Type II secretion: from structure to function. *FEMS Microbiol.Lett.*, 255:175-186.

Jongbloed, J.D., Antelmann, H., Hecker, M., Nijland, R., Bron, S., Airaksinen, U., et al. (2002) Selective contribution of the twin-arginine translocation pathway to protein secretion in Bacillus subtilis. *J.Biol.Chem.*, 277:44068-44078.

Käll, L., Krogh, A. and Sonnhammer, E.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J.Mol.Biol.*, 338:1027-1036.

Kandler, O. and Konig, H. (1978) Chemical composition of the peptidoglycan-free cell walls of methanogenic bacteria. *Arch.Microbiol.*, 118:141-152.

Knoblauch, N.T., Rudiger, S., Schonfeld, H.J., Driessen, A.J., Schneider-Mergener, J. and Bukau, B. (1999) Substrate specificity of the SecB chaperone. *J.Biol.Chem.*, 274:34219-34225.

König, H. (1988) Archaeobacterial cell envelopes. *Can. J. Microbiol.*, 34:395-395-406.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J.Mol.Biol.*, 305:567-580.

Kutsukake, K., Sukhan, A. and Yokoseki, T. (1994) Isolation and characterization of FliK-independent flagellation mutants from Salmonella typhimurium. *J. Bacteriol.*, 176:7625-7629.

Lee, H.C. and Bernstein, H.D. (2001) The targeting pathway of Escherichia coli presecretory and integral membrane proteins is specified by the hydrophobicity of the targeting signal. *Proc.Natl.Acad.Sci.U.S.A.*, 98:3471-3476.

Lescuyer, P., Hochstrasser, D.F. and Sanchez, J.C. (2004) Comprehensive proteome analysis by chromatographic protein prefractionation. *Electrophoresis*, 25:1125-1135.

Letoffe, S. and Wandersman, C. (1992) Secretion of CyaA-PrtB and HlyA-PrtB fusion proteins in Escherichia coli: involvement of the glycine-rich repeat domain of Erwinia chrysanthemi protease B. *J.Bacteriol.*, 174:4920-4927.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22:1658-1659.

Lloyd, S.A., Norman, M., Rosqvist, R. and Wolf-Watz, H. (2001) Yersinia YopE is targeted for type III secretion by N-terminal, not mRNA, signals. *Mol.Microbiol.*, 39:520-531.

Lower, M. and Schneider, G. (2009) Prediction of type III secretion signals in genomes of gram-negative bacteria. *PLoS One*, 4:e5917.

Lu, Z., Szafron, D., Greiner, R. Lu, P.,Wishart, D.S.,Poulin, B. *et al.* (2004) Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics*, 20:547-556.

Luirink, J., von Heijne, G., Houben, E. and de Gier, J.W. (2005) Biogenesis of inner membrane proteins in Escherichia coli. *Annu.Rev.Microbiol.*, 59:329-355.

Manisali, I., Chen, D.D., Schneider, B.B. (2006) Electrospray ionization source geometry for mass spectrometry: past, present and future. *TrAC. Trens. Analyt. Chem.* 25:243-256.

Mann, M. And Fenn, J.B. (1992) Electrospray mass spectrometry principle and method. In: Dedierio DM, Ed. Mass spectrometry: clinical and biomedical applications. New York: Plenum Press. 1-30.

Matsuda, S., Vert, J.P., Saigo, H., Ueda, N., Toh, H. and Akutsu, T. (2005) A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein Sci.*, 14:2804-2813.

Miao, E.A. and Miller, S.I. (2000) A conserved amino acid sequence directing intracellular type III secretion by Salmonella typhimurium. *Proc.Natl.Acad.Sci.U.S.A.*, 97:7539-7544.

Miyata, M. and Ogaki, H. (2006) Cytoskeleton of mollicutes. *J.Mol.Microbiol.Biotechnol.*, 11:256-264.

Moissl, C., Rachel, R., Briegel, A., Engelhardt, H. and Huber, R. (2005) The unique structure of archaeal 'hami', highly complex cell appendages with nano-grappling hooks. *Mol.Microbiol.*, 56:361-370.

Nagai, H., Cambronne, E.D., Kagan, J.C., Amor, J.C., Kahn, R.A. and Roy, C.R. (2005) A C-terminal translocation signal required for Dot/Icm-dependent delivery of the Legionella RalF protein to host cells. *Proc.Natl.Acad.Sci.U.S.A.*, 102:826-831.

Nair, R. and Rost, B. (2002) Sequence conserved for subcellular localization. *Protein Sci.*, 11:2836-2847.

Nakai, K. and Kanehisa, M. (1991) Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins*, 11:95-110.

Natale, P., Bruser, T. and Driessen, A.J. (2008) Sec- and Tat-mediated protein secretion across the bacterial cytoplasmic membrane--distinct translocases and mechanisms. *Biochim.Biophys.Acta*, 1778:1735-1756.

Ng, S.Y., Chaban, B. and Jarrell, K.F. (2006) Archaeal flagella, bacterial flagella and type IV pili: a comparison of genes and posttranslational modifications. *J.Mol.Microbiol.Biotechnol.*, 11:167-191.

Nickell, S., Hegerl, R., Baumeister, W. and Rachel, R. (2003) Pyrodictium cannulae enter the periplasmic space but do not enter the cytoplasm, as revealed by cryo-electron tomography. *J.Struct.Biol.*, 141:34-42.

Niehaus, F., Bertoldo, C., Kahler, M. and Antranikian, G. (1999) Extremophiles as a source of novel enzymes for industrial application. *Appl.Microbiol.Biotechnol.*, 51:711-729.

Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, 10:1-6.

Nielsen, H. and Krogh, A. (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc.Int.Conf.Intell.Syst.Mol.Biol.*, 6:122-130.

Niu, B., Jin, Y.H., Feng, K.Y., Lu, W.C., Cai, Y.D. and Li, G.Z. (2008) Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins. *Mol.Divers.*, 12:41-45.

Pace, N.R. (2006) Time for a change. *Nature*, 441:289.

Pallen, M.J.,Penn, C.W., and Chaudhuri, R.R. (2005) Bacterial flagellar diversity in the post-genomic era. *Trends Micriobiol.,* 13:143-149.

Patrie, S.M., Charlebois, J.P., Whipple, D., Kelleher, N.L., Hendrickson, C.L., Quinn, J.P., et al. (2004) Construction of a hybrid quadrupole/Fourier transform ion cyclotron resonance mass spectrometer for versatile MS/MS above 10 kDa. *J.Am.Soc.Mass Spectrom.*, 15:1099-1108.

Petersen, T.N., Brunak, S., von Heijne, G. and Nielsen, H. (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat.Methods*, 8:785-786.

Pohlner, J., Halter, R., Beyreuther, K. and Meyer, T.F. (1987) Gene structure and extracellular secretion of Neisseria gonorrhoeae IgA protease. *Nature*, 325:458-462.

Pohlschroder, M., Prinz, W.A., Hartmann, E. and Beckwith, J. (1997) Protein translocation in the three domains of life: variations on a theme. *Cell*, 91:563-566.

Pukatzki, S., Ma, A.T., Sturtevant, D., Krastins, B., Sarracino, D., Nelson, W.C., et al. (2006) Identification of a conserved bacterial protein secretion system in Vibrio cholerae using the Dictyostelium host model system *Proc.Natl.Acad.Sci.U.S.A.*, 103:1528-1533.

Ramamurthi, K.S. and Schneewind, O. (2002) Yersinia enterocolitica type III secretion: mutational analysis of the yopQ secretion signal. *J.Bacteriol.*, 184:3321-3328.

Rashid, M., Saha, S. and Raghava, G.P. (2007) Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. *BMC Bioinformatics*, 8:337.

Reinhardt, A. and Hubbard, T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.*, 26:2230-2236.

Restrepo-Montoya, D., Pino, C., Nino, L.F., Patarroyo, M.E. and Patarroyo, M.A. (2011) NClassG+: A classifier for non-classically secreted Gram-positive bacterial proteins. *BMC Bioinformatics*, 12:21.

Rey, S., Acab, M., Gardy, J.L., Laird, M.R., deFays, K., Lambert, C., et al. (2005a) PSORTdb: a protein subcellular localization database for bacteria. *Nucleic Acids Res.*, 33:D164-8.

Rey, S., Gardy, J.L. and Brinkman, F.S. (2005b) Assessing the precision of high-throughput computational and laboratory approaches for the genome-wide identification of protein subcellular localization in bacteria. *BMC Genomics*, 6:162.

Reysenbach, A.L., Liu, Y., Banta, A.B., Beveridge, T.J., Kirshtein, J.D., Schouten, S., et al. (2006) A ubiquitous thermoacidophilic archaeon from deep-sea hydrothermal vents. *Nature*, 442:444-447.

Robinson, C. and Bolhuis, A. (2004) Tat-dependent protein targeting in prokaryotes and chloroplasts. *Biochim.Biophys.Acta*, 1694:135-147.

Roggenkamp, A., Ackermann, N., Jacobi, C.A., Truelzsch, K., Hoffmann, H. and Heesemann, J. (2003) Molecular analysis of transport and oligomerization of the Yersinia enterocolitica adhesin YadA. *J.Bacteriol.*, 185:3735-3744.

Saltikov, C.W. and Newman, D.K. (2003) Genetic identification of a respiratory arsenate reductase. *Proc.Natl.Acad.Sci.U.S.A.*, 100:10983-10988.

Samudrala, R., Heffron, F. and McDermott, J.E. (2009) Accurate prediction of secreted substrates and identification of a conserved putative secretion signal for type III secretion systems. *PLoS Pathog.*, 5:e1000375.

She, S., Chen, F., Wang, K., Ester, M., Gardy, J.L. and Brinkman, F.S. (2003) Frequent-subsequence-based prediction of outer membrane proteins. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*:436-445.

Shen, H.B. and Chou, K.C. (2007) Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. *Protein Eng.Des.Sel.*, 20:39-46.

Shen, H.B. and Chou, K.C. (2009) Gpos-mPLoc: a top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins. *Protein Pept.Lett.*, 16:1478-1484.

Shen, H.B. and Chou, K.C. (2010) Gneg-mPLoc: a top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins. *J.Theor.Biol.*, 264:326-333.

Shetty, A., Chen, S., Tocheva, E.I., Jensen, G.J. and Hickey, W.J. (2011) Nanopods: a new bacterial structure and mechanism for deployment of outer membrane vesicles. *PLoS One*, 6:e20725.

Soler, N., Marguet, E., Verbavatz, J.M. and Forterre, P. (2008) Virus-like vesicles and extracellular DNA produced by hyperthermophilic archaea of the order Thermococcales. *Res.Microbiol.*, 159:390-399.

Solis, N. and Cordwell, S.J. (2011) Current methodologies for proteomics of bacterial surface-exposed and cell envelope proteins. *Proteomics*, 11:3169-3189.

Stanier, R.Y. and van Niel, C.B. (1962) The concept of a bacterium. *Arch.Mikrobiol.*, 42:17-35.

Stanley, P., Packman, L.C., Koronakis, V. and Hughes, C. (1994) Fatty acylation of two internal lysine residues required for the toxic activity of Escherichia coli hemolysin. *Science*, 266:1992-1996.

Su, E.C., Chiu, H.S., Lo, A., Hwang, J.K., Sung, T.Y. and Hsu, W.L. (2007a) Protein subcellular localization prediction based on compartment-specific features and structure conservation. *BMC Bioinformatics*, 8:330.

Sutcliffe, I.C. (2010) A phylum level perspective on bacterial cell envelope architecture. *Trends Microbiol.*, 18:464-470.

Thanassi, D.G. and Hultgren, S.J. (2000) Multiple pathways allow protein secretion across the bacterial outer membrane. *Curr.Opin.Cell Biol.*, 12:420-430.

Thoma, C., Frank, M., Rachel, R., Schmid, S., Nather, D., Wanner, G., et al. (2008) The Mth60 fimbriae of Methanothermobacter thermoautotrophicus are functional adhesins. *Environ.Microbiol.*, 10:2785-2795.

Thompson, B.G. and Murray, R.G. (1981) Isolation and characterization of the plasma membrane and the outer membrane of Deinococcus radiodurans strain *Sark. Can. J. Microbiol.,*27: 729-734.

Tjalsma, H., Bolhuis, A., Jongbloed, J.D., Bron, S. and van Dijl, J.M. (2000) Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome. *Microbiol.Mol.Biol.Rev.*, 64:515-547.

Troisfontaines, P. and Cornelis, G.R. (2005) Type III secretion: more systems than you think. *Physiology (Bethesda)*, 20:326-339.

Tusnady, G.E. and Simon, I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, 17:849-850.

Vergunst, A.C., van Lier, M.C., den Dulk-Ras, A., Stuve, T.A., Ouwehand, A. and Hooykaas, P.J. (2005) Positive charge is an important feature of the C-terminal transport signal of the VirB/D4-translocated proteins of Agrobacterium. *Proc.Natl.Acad.Sci.U.S.A.*, 102:832-837.

Viklund, H. and Elofsson, A. (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci.*, 13:1908-1917.

Vogel, U. and Claus, H. (2011) Vaccine development against Neisseria meningitidis. *Microb.Biotechnol.*, 4:20-31.

von Heijne, G. (1990) The signal peptide. *J.Membr.Biol.*, 115:195-201.

Voulhoux, R. and Tommassen, J. (2004) Omp85, an evolutionarily conserved bacterial protein involved in outer-membrane-protein assembly. *Res.Microbiol.*, 155:129-135.

Walther, T.C. and Mann, M. (2010) Mass spectrometry-based proteomics in cell biology. *J.Cell Biol.*, 190:491-500.

Wang, J., Sung, W.K., Krishnan, A. and Li, K.B. (2005) Protein subcellular localization prediction for Gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines. *BMC Bioinformatics*, 6:174.

Wang, Y., Zhang, Q., Sun, M.A. and Guo, D. (2011) High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. *Bioinformatics*, 27:777-784.

Wang, Z., Jiang, L., Li, M., Sun, L. and Lin, R. (2007) Fast Fourier transform-based support vector machine for subcellular localization prediction using different substitution models. *Acta Biochim.Biophys.Sin.(Shanghai)*, 39:715-721.

Weiner, J.H., Bilous, P.T., Shaw, G.M., Lubitz, S.P., Frost, L., Thomas, G.H., *et al.* (1998) A novel and ubiquitous system for membrane targeting and secretion of cofactor-containing proteins. *Cell*, 93:93-101.

Williams, A.W., Yamaguchi, S., Togashi, F., Aizawa, S.I., Kawagishi, I. and Macnab, R.M. (1996) Mutations in fliK and flhB affecting flagellar hook and filament assembly in Salmonella typhimurium. *J. Bacteriol.* 178:2960-2970.

Winsor, G.L., Van Rossum, T., Lo, R., Khaira, B., Whiteside, M.D., Hancock, R.E., *et al.* (2008) Pseudomonas Genome Database: facilitating user-friendly, comprehensive comparisons of microbial genomes. *Nucleic Acids Res.*

Woese, C.R. (2004) A new biology for a new century. *Microbiol.Mol.Biol.Rev.*, 68:173-186.

Woese, C.R. and Fox, G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc.Natl.Acad.Sci.U.S.A.*, 74:5088-5090.

Wulff-Strobel, C.R., Williams, A.W. and Straley, S.C. (2002) LcrQ and SycH function together at the Ysc type III secretion system in Yersinia pestis to impose a hierarchy of secretion. *Mol.Microbiol.*, 43:411-423.

Yu, C.S., Chen, Y.C., Lu, C.H. and Hwang, J.K. (2006) Prediction of protein subcellular localization. *Proteins*, 64:643-651.

Yu, L., Guo, Y., Li, Y., Li, G., Li, M., Luo, J. *et al.* (2010) SecretP: identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition. *J. Theor. Biol.*, 267:1-6.

Yu, N.Y., Laird, M.R.,Spencer, C. and Brinkman, F.S. (2011) PSORTdb - an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea. *Nucleic Acids Res,* 39(Database issue):D241-244.

Yu, N.Y., Wagner, J.R., Laird, M.R., Melli, G., Rey, S., Lo, R. *et al*. (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, 26:1608-1615.

Zhou, M., Boekhorst, J., Francke, C. and Siezen, R.J. (2008) LocateP: genome-scale subcellular-location predictor for bacterial proteins. *BMC Bioinformatics*, 9:173.

Zolghadr, B., Klingl, A., Rachel, R., Driessen, A.J. and Albers, S.V. (2011) The bindosome is a structural component of the Sulfolobus solfataricus cell envelope. *Extremophiles*, 15:235-244.