Creation, evaluation, and use of PSI, a program for identifying protein-phenotype relationships and comparing protein content in groups of organisms

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Brett Trost

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

Abstract

Recent advances in DNA sequencing technology have enabled entire genomes to be sequenced quickly and accurately, resulting in an exponential increase in the number of organisms whose genome sequences have been elucidated. While the genome sequence of a given organism represents an important starting point in understanding its physiology, the functions of the protein products of many genes are still unknown; as such, computational methods for studying protein function are becoming increasingly important. In addition, this wealth of genomic information has created an unprecedented opportunity to compare the protein content of different organisms; among other applications, this can enable us to improve taxonomic classifications, to develop more accurate diagnostic tests for identifying particular bacteria, and to better understand protein content relationships in both closely-related and distantly-related organisms.

This thesis describes the design, evaluation, and use of a program called Proteome Subtraction and Intersection (PSI) that uses an idea called genome subtraction for discovering protein-phenotype relationships and for characterizing differences in protein content in groups of organisms. PSI takes as input a set of proteomes, as well as a partitioning of that set into a subset of "included" proteomes and a subset of "excluded" proteomes. Using reciprocal BLAST hits, PSI finds orthologous relationships among all the proteins in the proteomes from the original set, and then finds groups of orthologous proteins containing at least one orthologue from each of the proteomes in the "included" subset, and none from any of the proteomes in the "excluded" subset.

PSI is first applied to finding protein-phenotype relationships. By identifying proteins that are present in all sequenced isolates of the genus *Lactobacillus*, but not in the related bacterium *Pediococcus pentosaceus*, proteins are discovered that are likely to be responsible for the difference in cell shape between the lactobacilli and *P. pentosaceus*. In addition, proteins are identified that may be responsible for resistance to the antibiotic gatifloxacin in some lactic acid bacteria.

This thesis also explores the use of PSI for comparing protein content in groups of organisms. Based on the idea of genome subtraction, a novel metric is proposed for comparing the difference in protein content between two organisms. This metric is then used to create a phylogenetic tree for a large set of bacteria, which to the author's knowledge represents the largest phylogenetic tree created to date using protein content. In addition, PSI is used to find the proteomic cohesiveness of isolates of several bacterial species in order to support or refute their current taxonomic classifications.

Overall, PSI is a versatile tool with many interesting applications, and should become more and more valuable as additional genomic information becomes available.

Acknowledgements

I would first like to thank my supervisor, Dr. Tony Kusalik, for his comments, advice, and support. Even more than these things, however, I appreciate most of all his always-plentiful enthusiasm for research. Special thanks to Monique Haakensen for sharing her microbiology knowledge, for thinking of interesting applications of PSI, and for assisting with biological interpretations. I would also like to thank Dr. Barry Ziola, Dr. Ian McQuillan, and Vanessa Pittet for their helpful comments and suggestions, Dr. Fiona Brinkman for agreeing to be the external examiner, and Dr. Raymond Spiteri for the use of dahlquist. Special thanks to my father, Randy, my mother, Ruth, and my sister, Kelli, for supporting and encouraging me throughout my education, and to Chantel, for making graduate studies a lot more fun than it would be otherwise. Funding for my Master's degree was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Department of Computer Science, and the College of Graduate Studies and Research.

Contents

List of Tables

List of Figures

List of Algorithms

List of Abbreviations

AUP	average unique proteins
BeT	best hit
BLAST	basic local alignment search tool
bp	base pair
cDNA	complementary DNA
COGs	clusters of orthologous groups
DFS	depth-first search
DNA	deoxyribonucleic acid
EBI	European Bioinformatics Institute
EDGAR	Efficient Database framework for comparative Genome Analyses using BLAST score Ratios
EMBL	European Molecular Biology Laboratory
EST	expressed sequence tag
GO	gene ontology
HMM	hidden Markov model
IMG	Integrated Microbial Genomes
LAB	lactic acid bacteria
MLSA	multi-locus sequence analysis
PANTHER	protein analysis through evolutionary relationships
PCR	polymerase chain reaction
RBH	reciprocal BLAST hits
RNA	ribonucleic acid
rRNA	ribosomal RNA
SQL	structured query language
UPGMA	unweighted pair group method with arithmetic mean
XML	extensible markup language

Chapter 1

Introduction

With the advent of modern sequencing techniques, the genomes of many organisms—from those with small genomes, like bacteria, to those with much larger genomes, such as plants—can be readily sequenced. As a result, hundreds of prokaryotic genomes have been sequenced to date, and more and more eukaryotic genomes are becoming available. Large quantities of sequence information provide the opportunity to perform interesting and useful comparisons of protein content in different organisms. The purpose of this thesis is to describe, evaluate, and apply a program called Proteome Subtraction and Intersection (PSI). PSI is based on an idea called "genome subtraction", which involves finding (using either computational or laboratory techniques) genetic regions that are present in some organisms, but not others. PSI's specific purpose is to find proteins that are present in all of the organisms in one set of organisms, and none of the organisms in a second set. The name of the program was chosen because doing this requires the subtraction and intersection set operations. Two primary applications of PSI are described below.

First, PSI can help answer the question, "What protein or proteins are responsible for phenotype X?" By finding proteins that are present in all of the organisms that exhibit phenotype X, but in none of the organisms that do not exhibit phenotype X, PSI can narrow down the list of proteins that could potentially cause this phenotype. Using this list, the amount of molecular biology laboratory work required to identify the precise protein or proteins responsible for X can be substantially reduced.

Second, PSI can provide novel insights concerning the protein content of groups of organisms. This is a very general statement of this application, and there are many more specific applications that fall under this category. For instance, PSI can be used to answer the question, "What proteins does species A contain that are found in none of the other organisms of the same genus as A?" If A is a bacterial species, then the resulting list of proteins may be useful for developing diagnostic procedures that can differentiate species A from similar types of bacteria, perhaps by testing for the presence or absence of the gene corresponding to one of these proteins. Such a gene could be detected using the polymerase chain reaction (PCR) laboratory technique. PSI can also generate data that is useful for performing whole-genome phylogenetic analyses by determining how many proteins are present in one organism, but not another. This quantity reflects a genomic "distance" between these two organisms, and presumably reflects an evolutionary distance as well. If all pairwise comparisons are performed among the organisms of interest, then a phylogenetic tree can be constructed using these distances. Such a phylogenetic tree takes into account much more information than traditional approaches to phylogenetics, which often just consider changes in one gene sequence or a small set of sequences. In addition to these questions, PSI can address many other problems that involve comparing the protein content in sets of organisms. Some of these issues will be addressed in this thesis, while others will be discussed as possible future work.

This thesis describes the design and implementation of PSI, and evaluates the efficacy of PSI for performing the aforementioned types of analyses. Given that much more sequence information is available for prokaryotes than for eukaryotes, the analyses presented in this thesis are restricted to prokaryotes (specifically bacteria). However, PSI could be applied to eukaryotes as well, although there might be added difficulties when using PSI to analyze these more complex organisms. Some of these potential difficulties are discussed in Section 6.3.

Background to the concepts presented in this thesis is given in Section 2. Section 3 summarizes the goals of this research. Section 4 describes the design and implementation of PSI, as well as the data and methods used to evaluate it. Section 5 presents the results. Finally, Section 6 gives some concluding remarks, discusses some issues relating to the results, and suggests possibilities for future work.

Chapter 2

Background

This section describes the background material necessary to understand the content of this thesis. Section 2.1 gives a basic introduction to molecular biology. Section 2.2 contains a short history of genome sequencing, as well as a discussion of the usefulness and the limitations of knowing an organism's genome sequence. Section 2.3 gives an introduction to online databases containing sequence information. Section 2.4 contains a short background on genetic mutations, while Section 2.5 discusses methods for determining the similarity of biological sequences. Section 2.6 reviews techniques for searching sequence databases. A survey of orthologue detection methods can be found in Section 2.7, and a short primer on the problem of determining protein-phenotype relationships is given in Section 2.8. Section 2.9 contains an introduction to phylogenetics. Finally, Section 2.10 gives background material on graphs, and Section 2.11 describes disjoint-set data structures.

2.1 Molecular biology

Molecular biology is the study of molecules that are important to life, specifically DNA (deoxyribonucleic acid), RNA (ribonucleic acid), and proteins. These three molecules are so inextricably linked, and their importance to biology so fundamental, that the relationship among them is called the central dogma of molecular biology. This section contains a short introduction to the structure and function of these three types of molecules, and describes how they are related.

DNA molecules are long polymers made up of the four chemical constituents adenine, cytosine, guanine, and thymine, which are abbreviated by the letters A, C, G, and T, respectively. Each of these chemicals is called a nucleotide, which is often used interchangeably with the word base. RNA is chemically very similar to DNA, and is also composed of repeating sequences of four building blocks; three of them (A, C, and G) are analogues to those in DNA, while RNA contains uracil (U) instead of thymine. Proteins are polymers consisting of sequences of 20 possible amino acid residues, and constitute both the building blocks and the machinery of a cell—some proteins have structural roles, while others perform the functions that a cell needs to survive, grow, and replicate.

DNA is the chemical that is passed down from generation to generation, and ultimately determines the characteristics of a given organism. DNA is used as a template for the synthesis of RNA;

in turn, RNA is (usually) used as a template for the synthesis of proteins. The central dogma of molecular biology summarizes very succinctly the relationship among these three types of molecules: information flows from DNA to RNA to proteins. It should be noted that in special cases, other paths of information flow are possible (RNA to DNA, for instance). Transcription is the process by which RNA is synthesized using DNA as a template, while translation is the process by which proteins are made using RNA as a template. The entire DNA complement of an organism is called its genome, and the complement of proteins synthesized by an organism is called its proteome.

A gene is a segment of DNA that contains all of the information necessary for the synthesis of an RNA molecule, and can be considered the basic unit of heredity [1]. The term genotype can have slightly different meanings depending on the particular context, but usually refers to the presence or absence of one or more genes, or the specific form of a given gene that is present in a particular organism. Very generally, a genotype can be described as a genetic property of an organism. A phenotype is an actual observable characteristic produced by a genotype; for example, a particular bacterium might have a rod-like cell shape, while another may have a round cell shape. This difference in cell shape (the phenotype) may be caused by differences in genotype—for instance, one bacterium may contain a gene not found in the other; alternatively, the difference could be due to the two bacteria expressing a different form (allele) of the same gene.

2.2 Genome sequencing

In 1977, Fred Sanger and colleagues used the plus and minus method [2] to determine the complete genome sequence of bacteriophage ϕX174 [3]. Consisting of fewer than 6000 base pairs (bp), the ϕX174 genome was the first complete genome sequence to be reported. At this time, sequencing technology was rather primitive, making sequencing efforts expensive, laborious, and error-prone. As a result, only very short genomes—those with lengths similar in magnitude to that of bacteriophage ϕX174—could be sequenced in a reasonable amount of time. With the advent of more advanced sequencing techniques, such as the dideoxyribonucleotide chain-termination sequencing method [4], DNA sequencing could be performed at a more rapid pace. This made it possible to elucidate the sequences of several larger viral genomes, such as those of bacteriophage λ (which consisted of 48502 bp) [5], vaccinia virus (191636 bp) [6], and variola (smallpox) virus (186102 bp) [7].

While the sequencing of these larger viral genomes was an impressive achievement, the limitations of then-current sequencing technology put the genome sequences of non-viral organisms largely out of reach. Even the smallest prokaryotic genomes are several times the size of the largest viral genomes, and eukaryotic genomes can be orders of magnitude larger than that. However, new and improved techniques for genome sequencing were soon introduced [8–11]. These methods enabled larger and larger genomes to be sequenced in a reasonable amount of time, and soon led to several important milestones in genome sequencing. In 1995, the genome sequence of the bacterium *Haemophilus influenzae* was reported, which represented the first published sequence of the genome of a free-living organism [12]. This genome sequence consisted of nearly two million base pairs—far larger than any of the viral genomes that had previously been reported. The sequencing of the yeast *Saccharomyces cerevisiae* [13] in 1996 represented the first genome sequence of a eukaryotic organism. Another important genome sequence was that of the animal model organism *Drosophila melanogaster* [14], which was followed closely by the genome sequence of the plant model organism *Arabidopsis thaliana* [15]. Completed in 2001, the sequencing of the human genome [16, 17] was perhaps the biggest milestone of all in terms of human medical significance.

These sequencing efforts were very significant achievements. However, an organism's genome sequence does not tell us what genes are present, where those genes are physically located (such as on chromosomes or plasmids), or the functions of the protein products of those genes. Thus, knowing an organism's genome sequence represents only a starting point in understanding its genetic properties. Much work still needs to be done in order to enhance our understanding of the instructions that guide the organism's development, dictate the metabolic processes that it can perform, and allow it to respond to its environment. It is difficult to overstate the value of gaining such knowledge; among many other important applications, it could allow us to improve human health (say, by discovering how viruses are able to evade the host's immune response), to improve agricultural yields (say, by engineering herbicide-resistant crops), and to facilitate industrial processes (say, by using bacterial enzymes to catalyze chemical reactions).

2.3 Databases of genome and proteome sequences

Due to exponential increases in the number of sequenced genomes, central repositories are needed so that these sequences can be easily organized and accessed. Part of the European Bioinformatics Institute (EBI) website, Integr8 [18] is an online database that contains genome and proteome sequences. As of June 4, 2009, Integr8 had available for download the genomes and proteomes for 830 bacterial isolates [19], 63 archaeal isolates [20], 1930 viral strains [21], and 96 eukaryotes [22]. Sequences are available for download in a number of formats—genome sequences can be retrieved in FASTA or European Molecular Biology Laboratory (EMBL) format, while proteomes are available in FASTA, UniProt, or extensible markup language (XML) format.

2.4 Genetic mutations

During evolution, many changes may occur to an organism's genome due to errors in DNA replication. For instance, a single nucleotide may be substituted for another nucleotide; alternatively, a nucleotide may be added (an insertion) or omitted (a deletion). More large-scale changes also can take place. For example, a translocation occurs when an entire segment of a chromosome is removed and re-inserted into a different position in the same chromosome, or (in eukaryotes) into a different chromosome. A reversal occurs when a segment of a chromosome is removed, and then reinserted in the original place, but in reverse orientation. Another possible genetic change is gene duplication, in which duplication of a gene occurs. After a gene duplication event, the original gene often retains its original function, while the copy evolves to perform a new, but usually related, function.

2.5 Similarity of DNA and protein sequences

For a given protein or DNA sequence from a specific organism, it is frequently of biological interest to find other sequences from the same organism or from a different organism that are similar to it. Two genes or proteins that are similar are generally assumed to be evolutionarily related. To describe more precisely how two sequences may be related in an evolutionary sense, it is helpful to introduce some terminology. Note that the following descriptions discuss similarity among proteins, but are also valid when considering DNA sequences. A given protein is a homologue of another protein if the two proteins are evolutionarily related. A paralogue is a more specific version of a homologue, and refers to a protein that is related to another protein from the same organism by virtue of a gene duplication event [23]. An orthologue is also a more specific version of a homologue, and refers to a protein that is related to another protein from a different organism by virtue of both proteins having evolved from a single ancestral protein. Orthology is a concept that is central to this thesis.

An example of orthology is as follows. Suppose that the genome of organism A encodes protein X_A. Over time, the descendants of organism A diverge, eventually differentiating into two species B and C. During this differentiation process, protein X_A undergoes mutations, and the versions of this protein present in organisms B and C can be denoted X_B and X_C, respectively. Despite these mutations, X_B and X_C retain the same function as X_A. As such, X_B and X_C are orthologues.

The most basic method for ascertaining the similarity of two protein sequences is to calculate the optimal alignment of those sequences, and then determine the proportion of amino acid residues that are either identical, or have similar biochemical properties. The alignment portion of this procedure could involve either a global sequence alignment, in which the entirety of the two sequences are aligned, or a local sequence alignment, in which only the most similar portions of the sequences are aligned. The algorithms used to perform global alignment and local alignment both use an algorithmic technique called dynamic programming, and were introduced by Needleman and Wunsch [24] and Smith and Waterman [25], respectively.

2.6 Database searching

While mathematically optimal, the dynamic programming method of aligning sequences is too slow to be practical for comparing a sequence against a large database of other sequences. For this reason, faster heuristic-based methods are required. Heuristic-based methods are not mathematically optimal, but are much faster and often give satisfactory results. The first widely-used programs used for searching sequence databases were the FASTP program [26] and its successor, FASTA [27]. Note that the FASTA program for sequence database searching should not be confused with the FASTA file format for representing sequence information. The FASTP and FASTA programs allow large databases to be searched using ordinary computers in seconds or minutes, compared to hours or days for the dynamic programming methods.

Today, the most widely used database search tool is BLAST ("basic local alignment search tool") [28, 29], which provides greater specificity and similar sensitivity when compared with the FASTA algorithm, and is also much faster. For protein sequences, BLAST first uses a sliding window to find all the words (sequences of characters) of length three in the input sequence. For each word w, a similarity matrix is used to find other three-letter words that are similar to w. For each original word (from the query sequence), as well as the words that are similar to them, the database is searched for that word. Each time a word is found, a "seed" is created, which is extended in both directions. Put differently, an alignment is created between the query sequence and the sequence in the database, originating at the seed and extending outward. During this extension, the current score of the alignment is kept, as well as the maximum score achieved thus far. The extension is terminated when the current score drops below the maximum score by a certain amount, and the alignment that gave the maximum score is reported. BLAST outputs statistical measures of significance for each match in the database, aiding the user in determining whether a given match has biological meaning. The most commonly used statistical measure of significance reported by BLAST is the E-value. The E-value represents, for a given sequence with score S, the expected number of matches obtaining a score equal to or better than S that would occur by chance given the size of the database. The smaller the E-value, the smaller the chance that the match occurred simply by chance.

2.7 Orthologue detection

Several techniques have been proposed for identifying orthologous relationships among proteins. As orthologue detection is a concept that is central to this thesis, these methods are explained in some detail. Section 2.7.1 discusses a simple method for orthologue detection, while Section 2.7.2 describes a slightly more sophisticated one. Section 2.7.3 discusses clusters of orthologous groups

(COGs), a popular orthologue database. Finally, Section 2.7.4 contains a brief overview of some additional orthologue detection methods.

2.7.1 A simple method for orthologue detection

Perhaps the simplest possible approach to orthologue detection is to perform pairwise BLAST searches between every possible pair of proteins, and to declare two proteins orthologues if one (or both) of the matches has an E-value that is less than some threshold. Unfortunately, this method is prone to identifying two proteins as orthologues even when they are not. To understand why, consider the following hypothetical example.

The genomes of two closely related bacteria, O_1 and O_2, each encode a protein (P_1 and P_2, respectively) that allows them to metabolize glucose. At some point, a gene duplication event occurs in O_1, and a duplicate gene encoding P_1 is now present in O_1's genome. Denote these proteins P_1^a and P_1^b. Over time, P_1^a retains its original function, while P_1^b evolves to obtain a new function. Now, P_1^a and P_1^b would be termed paralogues, while P_1^a and P_2 would be termed orthologues. However, P_1^b and P_2 are not orthologues, since they have different functions. Now consider the situation in which the relationships among these three proteins are not known in advance. If BLAST was simply run using P_1^b as the query sequence against the database of proteins encoded by the genome of O_2, then P_2 would likely be a very significant match (i.e., having a small E-value), leading to the erroneous conclusion that P_1^b and P_2 are orthologues.

2.7.2 Reciprocal BLAST hits

An improvement over the orthologue detection method described in Section 2.7.1 is called reciprocal BLAST hits (RBH). The principle behind RBH is simple: two proteins P_1 and P_2 (from organisms O_1 and O_2, respectively) are considered to be orthologues if and only if the following criteria are met.

- P_2 is the best hit (i.e., having the smallest E-value) when P_1 is used as the query sequence and the proteins in O_2 are used as the database.
- P_1 is the best hit when P_2 is used as the query sequence and the proteins in O_1 are used as the database.
- The E-values reported for both comparisons are each less than some threshold.

Continuing with the example from Section 2.7.1, the following shows the RBH would correctly determine the relationships among all three proteins. P_1^a would be the best BLAST hit when P_2 is used as the query sequence against the database of proteins in O_1, and P_2 would be the best BLAST hit when P_1^a is used as the query sequence against the database of proteins in O_2. Thus, P_1^a

and P_2 would be correctly identified as orthologues. However, while P_2 would be the best BLAST hit when P_1^b is used as the query sequence against the database of proteins in O_2, P_1^b would not be the best BLAST hit when P_2 is used as the query sequence against the database of proteins in O_1. Thus, unlike the simpler method described in Section 2.7.1, RBH would not incorrectly identify P_1^b and P_2 as orthologues.

2.7.3 Clusters of orthologous groups

One of the first significant attempts to study orthology among proteins from sequenced genomes was performed by Tatusov et al. [23], who introduced the concept of COGs. At the time that the first paper on COGs was published, the genomes for only seven free-living organisms (i.e., excluding viruses) were available, although all five major phylogenetic lineages (gram-negative bacteria, gram-positive bacteria, cyanobacteria, archaea, and eukaryotes) were represented.

The procedure used to construct the COGs was fairly simple: first, all possible pairwise BLAST searches were performed between proteins from the seven organisms. For each protein from a given organism, the best BLAST hit from each of the other six organisms was determined. These results were encoded using a graph; vertices represented proteins, and directed edges were drawn from one vertex to another if the second vertex was the best BLAST hit when the first was used as a query sequence. For a more detailed description of graphs, see Section 2.10. Triangles in the graph—where there was an edge between protein A and protein B, an edge between protein B and protein C, and an edge between protein C and protein A—represented the smallest possible COGs. The directionality of each of these edges was not important—a given edge could be going in either direction (from protein A to protein B, or from protein B to protein A), or could be bidirectional (both from protein A to protein B, and from protein B to protein A). Visual representations of COGs do indicate the directionality of each edge, even though the directionality was not important in constructing the COGs. It is also possible that cycles (other than triangles) could arise in the graph. These were not considered COGs; a minimal COG was created only if a triangle pattern occurred. After identifying triangles, these minimal COGs were then expanded by locating pairs of COGs whose triangles shared a common side, and joining them. This process was repeated until no more COGs could be joined together. A visual representation of COGs is shown in Figure 2.1.

While some of the procedure for constructing COGs was automated, a large amount of manual curation was required [31]. In particular, a specific COG created using the automated procedure may contain more than one protein from the same species (as in Figure 2.1 (C)), and thus splitting up the COG may more accurately reflect orthologous relationships. COGs needing to be split up were identified by manually inspecting the sequences or by performing multiple sequence alignments. In the opposite situation, manual intervention was sometimes required to merge two COGs that appeared to contain proteins that evolved from a single ancestral protein. A number of other

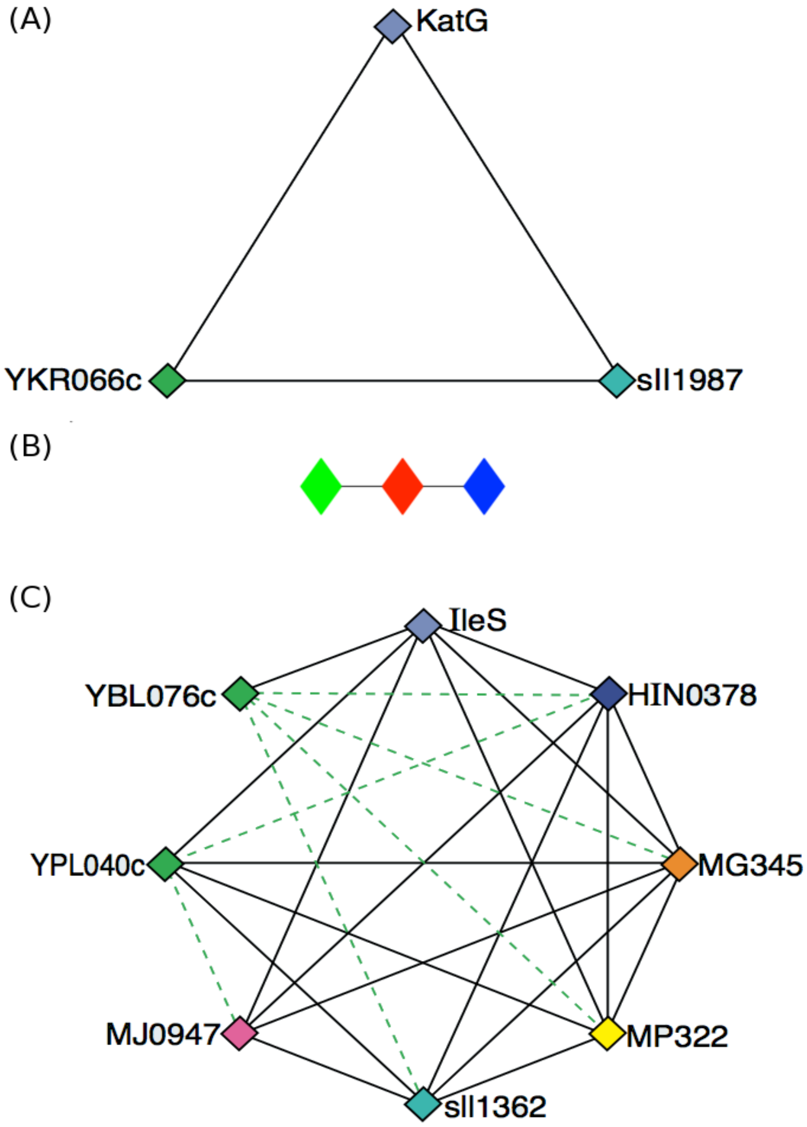

Figure 2.1: A visual representation of clusters of orthologous groups (COGs). Vertices representing proteins from the same species are the same color. Solid lines indicate bidirectional edges, in which the first protein is the best hit when the second is the query sequence, and vice versa. Broken lines indicate unidirectional edges, and the color of such an edge is the same as that of the node corresponding to the query sequence. Note that this notation is different from standard graph notation, in which arrows are used to indicate directionality. Part (A) shows the smallest possible COG, which consists of three proteins arranged in a triangle shape. This triangle shape is required for these proteins to be classified as a COG; three proteins for which the similarity relationship is not transitive, as shown in part (B), would not constitute a COG. Part (C) shows a more complex COG, with proteins from a few different species, and two proteins (which are paralogues) from the same species (YBL076c and YPL040c). Parts (A) and (C) were taken from Tatusov et al. [23], while part (B) was created using GraphViz [30].

manual curation procedures were performed in order to identify proteins missing from COGs and to support or refute the orthology of certain groups of proteins.

Following the publication of the original paper on COGs, the COG database was subsequently updated to include information from 21 complete genomes [32]. Even more recent updates have incorporated information from additional genomes [33]. Another paper describes the creation of clusters consisting exclusively of proteins from eukaryotes [34].

COGs have also been constructed from smaller sets of bacteria in order to compare their genomic content. For instance, Makarova et al. [35] sequenced the genomes of nine different lactic acid bacteria (LAB), which are bacteria characterized by their ability to ferment hexose sugars to form lactic acid. Makarova et al. created COGs containing proteins only from these organisms, which they called LaCOGs. They then compared the LaCOGs to previously-established COGs in order to characterize the genomic content of these newly-sequenced species, and to determine how the genomes of these LAB are related to each other and to the genomes of other bacteria. Another example is arCOGs, which are clusters of orthologous groups for 41 archaeal genomes [36].

While COGs represent an important tool for understanding orthologous relationships among proteins, they have a number of drawbacks that limit their usefulness for some applications. These are as follows.

1. In creating the COGs, two proteins P_1 and P_2 (from organisms O_1 and O_2, respectively) were connected with an edge if at least one of the following was true.

   (a) P_2 was the best hit (BeT, using the authors' terminology) when the query sequence P_1 was used to search the database of proteins encoded by O_2 using BLAST.

   (b) P_1 was the BeT when the query sequence P_2 was used to search the database of proteins encoded by O_1.

   Thus, two proteins were connected with an edge if one was the BeT of the other, regardless of the E-value of that hit. This could easily lead to spurious matches. For instance, suppose that P_2 was the BeT when P_1 was used as the query sequence and the proteins encoded by O_2 were used as the database. This may lead one to believe that P_1 and P_2 are orthologous. However, if the E-value for this match was relatively large—say, greater than 10^{-2}—it is possible that the two proteins were not actually related. To further illustrate this deficiency, suppose there exist two sets of random proteins. If an arbitrary protein from one set was used as a BLAST query against the other set, there must be a "best hit", even if all of the proteins in the two sets were entirely unrelated. If the sets were big enough, it is entirely plausible that this would result in some spurious triangles, which would be designated as COGs. Therefore, the use of BeTs for establishing the relatedness of proteins, and the corresponding lack of

E-value thresholds, could have created spurious COGs, or could have erroneously introduced a given protein into an existing COG.

Furthermore, note that only one of the two conditions above must be true in order for an edge to be drawn between two vertices. This is in contrast to the RBH method described in Section 2.7.2. Thus, COGs could potentially suffer from the same problem as the method described in Section 2.7.1. However, the fact that triangles were required to form a COG should have compensated for the fact that the BLAST hits were not required to be bidirectional. This, combined with the fact that the COGs were manually refined, suggests that COGs should certainly be superior to the simple method described in Section 2.7.1.

2. The composition of the COGs is dependent on the order in which genomes were added to the COGs [33]. Thus, if the 66 organisms whose proteins have been incorporated into COGs had been added in a different order, then the total number of COGs, as well as the proteins that were included in each of those COGs, could be different. While the authors analyzed how the number of COGs varied depending on the order in which genomes were added [33], they did not investigate how the actual composition of the COGs differed. If there was an appreciable degree of variation, the status of a given COG as a reliable reflection of protein orthology might be called into question.

   Given the procedure for constructing COGs, it is unclear why the order in which the genomes were added changes the number and composition of the COGs. Suppose that some genomes were added using the automatic procedure (finding triangles in the graph and then joining them), and then some manual curation was done, and then more genomes were added using the automatic procedure, followed by more manual curation, and so on. In this case, it makes sense that the number and composition of the COGs may differ depending on the order in which the genomes were added. However, in their 2001 paper [33], Tatusov et al. tried 10^6 possible genome order permutations, and presented a graph showing the range in the number of COGs that existed depending on the order in which the genomes were added. Clearly, with so many permutations being examined, no manual curation could have been done. Given that the automatic procedure for constructing COGs merely involved finding triangles and joining those with a common side, it is unclear why the order would matter. It is also unclear why all of the genomes could not be added at the same time.

3. While the initial construction of COGs was automated, a large amount of manual curation was necessary [34]. Since genomes are being sequenced at an ever-increasing rate, continuing this manual curation process becomes less and less feasible as more and more genome sequences become available.

4. In 2003, when the most recent paper discussing updates to the COG database was pub-

lished [34], the database contained 66 genomes. Despite the huge number of genomes that have been sequenced since that time (see Section 2.3 for specific numbers), the COG website still contained only 66 genomes as of Dec. 3, 2008 [37]. This is likely due in large part to the weakness discussed in item #3—namely, the significant amount of manual labour necessary to curate and refine COGs. The COG website claims that a large number of microbial genomes (261) and a few eukaryotic genomes are due to be added. However, even these numbers pale in comparison to the number of genomes that have been sequenced to date, and it is unclear when the integration of these genomes into the COG database will actually be completed.

5. As another consequence of this manual curation process, it is essentially impossible for those using COGs to determine how a particular COG was created. It is possible that a given COG was constructed entirely using the automatic process; alternatively, it may have been the result of an unknown amount of manual curation. As such, it is difficult to understand a given COG's relationship to related COGs, as well as to evaluate how well a given COG may reflect actual orthologous relationships.

6. The percentage of proteins from a given organism that are represented in COGs is always quite a bit less than 100%, making COGs inappropriate for some comparative genomics applications in which all, or the vast majority, of the proteins from the organisms of interest need to be considered. Table 2.1 shows the percentage of proteins that are represented in COGs for ten prokaryotic and ten eukaryotic organisms. Considering just the organisms in Table 2.1, the highest percentage of an organism's proteome represented in COGs is 73%, which is the case for four of the listed prokaryotes. Even for organisms with relatively small proteomes, such as the lactic acid bacterium *Lactobacillus acidophilus* (1864 proteins), the raw number of proteins not represented in COGs can be substantial (431 in this case). Table 2.1 shows that eukaryotes generally have a smaller percentage of proteins represented in COGs than do prokaryotes—if prokaryotes are considered in aggregate (1211 organisms, including both bacteria and archaea), 69% of proteins are represented in COGs; the corresponding percentage for eukaryotes (40 organisms) is just 33%. This lack of coverage makes COGs ill-suited for analyses in which a complete comparison of the proteins in a set of organisms is required. Note that all of the data in this paragraph were derived from the Integrated Microbial Genomes (IMG) website [38–40].

   As a side note, it is rather perplexing how IMG lists the percentage of proteins represented in COGs for all 1284 (as of July 20, 2009) prokaryotes included in its database, given that the COG database contains only 66 genomes. One possible explanation is that these values were derived using the COGNITOR program, which is available at the COG website. The COGNITOR program allows one to include an arbitrary protein P in an existing COG. If

Table 2.1: Percentage of proteins from various prokaryotic and eukaryotic organisms that are represented in COGs. Organisms were arbitrarily selected for inclusion in this table. All information was obtained from the Integrated Microbial Genomes website [38–40].

Organism	**Proteins (#)**	**Proteins represented in COGs (%)**
	Prokaryotes	
Bacteroides coprocola M16, DSM 17136	4291	47
Campylobacter hominis ATCC BAA-381	1741	62
Ignicoccus hospitalis KIN4/I	1444	65
Lactobacillus acidophilus NCFM	1864	73
Mycobacterium bovis AF2122/97	3949	69
Neisseria gonorrhoeae NCCP11945	2674	58
Oceanibulbus indolifex HEL-45	4153	73
Pyrococcus abyssi GE5	1904	73
Vibrio cholerae 1587	3758	73
Xanthomonas campestris pv. campestris	4273	69
	Eukaryotes	
Arabidopsis thaliana	26735	36
Aspergillus niger	14086	39
Caenorhabditis elegans	20056	25
Danio rerio	37724	30
Drosophila melanogaster	14081	25
Gibberella zeae	11640	46
Homo sapiens	27727	22
Magnaporthe grisea	12832	36
Mus musculus	39625	15
Trypanosoma brucei	8772	29

two or more proteins from different proteomes are represented in a given COG, and they are each the best BLAST hit for P in those proteomes, then P is included in that COG. Thus, the percentages listed on the IMG website could have been obtained by performing this procedure with each protein in a given organism (if this organism was not one of the 66 included in the COG database), and determining the percentage of these proteins that fit into an existing COG. Note that the purpose of this paragraph is to attempt to resolve an apparent contradiction between the number of organisms currently represented in the COG database, and the number of organisms for which IMG displays COG statistics. The documentation at IMG does not actually state that the COGNITOR program was used to generate these data, but this explanation seems plausible.

7. The COG website [37] has an unintuitive interface that is very difficult to use. In addition, the online documentation is sparse and confusing, and there are no tutorials or examples that allow the user to learn how to effectively obtain useful information from the COG database. Furthermore, the website makes extensive use of abbreviations and one-letter codes, and the same code is often used to refer to two things with entirely different meanings. For instance, the website uses the letter "P" to refer to both a category of protein function ("inorganic ion transport and metabolism"), and a specific clade (consisting of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*).

2.7.4 Other methods for orthologue detection

Besides the methods described in Sections 2.7.1, 2.7.2, and 2.7.3, a number of other techniques for the identification of orthologues have been developed. This section briefly describes some of these.

A common problem encountered in orthologue identification is that gene loss among closely related organisms may lead to two proteins being incorrectly classified as orthologues, when in fact they are paralogues. For instance, suppose that organism O_1's genome encodes protein P_1, and a gene duplication event occurs, resulting in two proteins denoted P_1^a (which retains its original function) and P_1^b (which gains a new function) in the genome of O_1. Now suppose that O_1's distant descendants, O_2 and O_3, encode proteins P_2^a and P_3^a, respectively (which are both orthologous to P_1^a), as well as P_2^b and P_3^b (which are both orthologous to P_1^b). If, due to gene loss, organism O_2 loses protein P_2^a and organism O_3 loses protein P_3^b, then the remaining proteins—P_2^b and P_3^a—would be incorrectly identified as orthologues by most orthologue detection techniques (such as RBH). The Ortholuge method [41] attempts to address this issue by analyzing previously-predicted orthologues for cases in which paralogues have been incorrectly identified as orthologues.

Another technique called OrthologID [42] uses a phylogenetic and parsimony-based approach to identifying orthologues. While previous work had described methods for identifying orthologues based on the manual inspection of trees, OrthologID was the first technique that enabled this procedure to be fully automated.

In addition to the orthologue detection techniques described above, other methods include RIO [43], Orthostrapper [44], and INPARANOID [45, 46].

A comparison of several orthologue detection techniques was performed by Chen et al. [47], who divided the different techniques into two groups: those that primarily use BLAST, and those that primarily use phylogenetic/tree-based techniques. Their paper reports that the BLAST-based methods and the tree-based methods each had characteristic strengths and weaknesses, with BLAST-based methods tending to be more sensitive and tree-based methods tending to be more specific. Thus, the choice of tool may depend on the relative importance of sensitivity versus specificity for a user's particular application. Of particular note is the authors' evaluation of the concordance of the orthologue classifications made by the various tools. Substantial disagreements in the classifications made by the various techniques were discovered, which may be partly due to the absence of a "gold standard" set of orthologues to which the output of a given tool can be compared.

2.8 Determining protein-phenotype relationships

The wealth of information provided by genome sequencing efforts has contributed greatly to elucidating the functions of the protein products of many genes. However, the genetic program encoded by an organism's genome is extremely complex, and even with the full genome sequences of hundreds of species, much remains unknown about the regulation, function, and physiological significance of many proteins. For instance, there are 4354 proteins in the human proteome (downloaded from UniProtKB [18] on Apr. 30, 2007) that contain in their annotations one or more of the words "putative", "uncharacterized", or "like". (An example of a protein annotation containing "like" is "Actin-like protein 6A".) These words are generally used to describe proteins whose functions are either completely unknown, or which can be only tentatively assigned based on similarity to other proteins. As the aforementioned UniProtKB set contains 38009 proteins, uncharacterized proteins constitute more than 10% of this set. The actual proportion of uncharacterized proteins is probably greater than this, as some proteins with unknown functions may not contain any of these "uncertainty words" in their annotations. When interpreting these data, it is important to note that humans are one of the most intensively studied species (albeit also one of the most complex), so one might expect other sequenced organisms to have an even greater deficit of protein characterization. Indeed, this appears to be the case: the proteome of *Streptococcus pyogenes* serotype M3, for instance, contains 1852 proteins, and 1387 of these are annotated as "putative", "uncharacterized", and/or "like". For the better-characterized bacterium *L. acidophilus* NCFM, 752 out of 1859 proteins are uncharacterized based on this criterion. Interestingly, even the extremely well-studied organism *Escherichia coli* K12 contains 99 uncharacterized proteins out of 3990 proteins in its proteome.

Besides determining the function of a specific protein, there is also the complementary problem of determining, for a given phenotype, the protein or proteins responsible for causing that phenotype. There are two main biological techniques for determining whether protein X indeed creates phenotype Y. First, gene knock-out organisms can be created, in which an unrelated piece of DNA is inserted in the middle of the gene encoding protein X, rendering its protein product nonfunctional. If phenotype Y is no longer exhibited by that organism, then it becomes clear that protein X causes that phenotype. In many cases, the situation is more complex, and X may be part of an entire pathway of catalysis or protein–protein interactions that result in phenotype Y. In this case, preventing the expression of X would still disrupt the phenotype, although it would be incorrect to conclude that the phenotype is solely due to the presence of X. Second, it is possible to create a gene knock-in organism, in which the gene encoding X is added to the genome of an organism that normally does not contain it. If this organism acquires phenotype Y, then one can conclude that protein X was responsible. However, if the organism does not acquire Y, then it does not necessarily mean that X has nothing to do with Y—it may be the case that X is just one of many proteins in a pathway responsible for Y, and the gene knock-in organism may not have the rest of the elements of the pathway necessary to cause this phenotype.

2.9 Phylogenetics

Phylogenetics is the study of the evolutionary relatedness of different organisms. This section outlines different techniques for studying phylogenetics. Specifically, Section 2.9.1 outlines both traditional and modern approaches to phylogenetics, and Section 2.9.2 presents a survey of whole-genome approaches.

2.9.1 Methods for studying phylogenetics

Historically, phylogenetic analyses have been performed using a diverse and often arbitrary selection of morphological and phenotypic characteristics. For instance, a newly-isolated bacterium would perhaps have been classified into one genus if the cells were round, and another genus if the cells were rod-shaped. However, it is now considered doubtful that individual phenotypes—or even a small collection of phenotypes—can be used to accurately infer evolutionary relationships [48]. This has led to the development of more reliable and accurate approaches for studying phylogenetics and classifying organisms.

Most modern phylogenetic analyses deduce evolutionary relationships using biomolecular sequences. The most popular approach to bacterial phylogenetics involves comparing 16S ribosomal

RNA (rRNA) gene sequences [48]. By comparing the similarity between the 16S rRNA gene sequences of two bacteria, their degree of evolutionary relatedness can be inferred. The 16S rRNA gene is part of the prokaryotic ribosome, and has a number of qualities that make it ideally suited for phylogenetic analysis: it is present in all prokaryotes; its function is always the same and is not involved in environmental response, meaning that the gene does not experience different evolutionary pressures depending on the environment in which the bacterium lives; it is easy to sequence; and different regions mutate at different rates, enabling both close and distant phylogenetic relationships to be analyzed. The 16S rRNA gene is present only in bacteria; however, portions of the eukaryotic ribosome that are similar to it are often used when studying eukaryotes.

Another common technique is multi-locus sequence analysis (MLSA) [49, 50] which infers phylogenetic relationships by comparing the sequences of several housekeeping genes. Since this technique takes several genes into account, rather than just a single gene, it is potentially more accurate.

Several software tools can be downloaded in order to visualize phylogenetic trees. A website containing a list of these programs is `http://bioinfo.unice.fr/biodiv/Tree_editors.html`. There also exist web-based tools that enable phylogenetic trees to be manipulated; a prominent example is the interactive tree of life (iTOL) website [51].

2.9.2 Whole genome approaches to phylogenetics

While 16S rRNA gene sequence analysis and MLSA have proved to be effective tools for phylogenetics, one deficiency inherent in these techniques is that the amount of information used is quite small relative to the total amount of information present in an organism's genome. As a result, there has been increasing interest in using whole-genome characteristics in analyzing evolutionary relationships.

One prominent example of a whole-genome similarity measure is the frequency of each possible dinucleotide. These frequencies have been found to be similar in closely related organisms and dissimilar in more distantly related organisms, and therefore constitute a “genomic signature” [52, and references therein]. Even before many genomes were available, dinucleotide frequencies in different organisms were characterized and compared using the sequence data available at the time [52]. More recently, van Passel et al. [53] evaluated the use of this genome signature for phylogenetics using a large number of prokaryotic genome sequences. Using a calculation called δ^*, which represents the average difference in abundance for all dinucleotides in two genomes [54], they showed that intraspecific distances are generally much smaller than interspecific (but intrageneric) distances. For a given pair of organisms, they also observed an inverse relationship between δ^* and the percent identity of their 16S rRNA genes, although the strength of this relationship appeared to be quite modest and in fact was not precisely quantified by the authors.

Many other whole-genomic approaches to taxonomy have been explored. A genome's G-C content (the percentage that is composed of guanine or cytosine) has been found to be highly similar in related species and less similar in more divergent species [55]. Similar patterns have been discovered for codon usage [56, 57] and gene order [58]. A particularly interesting approach to genomic phylogeny was introduced by Qi et al. [59, 60], who developed a program called CVTree that ascertains phylogenetic distances by examining the short peptide composition of each proteome. These methods, as well as a number of others, were reviewed by Coenye et al. [61].

Coenye and Vandamme [57] performed a comparison of some of these methods, and showed that the phylogenetic trees derived from these characteristics are usually quite consistent with each other, as well as with the tree derived from comparing 16S rRNA gene sequences. As these comparisons were performed on a small, related group of bacteria, it remains unclear whether these results generalize to all organisms or even to all bacteria.

Another approach to whole-genome phylogenetics—one that was also examined in the study done by Coenye and colleagues—is the comparison of gene content, which involves identifying orthologues in pairs of organisms, and then assigning a "distance" between each pair based on the number of shared genes. This technique was originally proposed by Snel et al. [62], and has subsequently been revisited with larger groups of organisms [63, 64]. Compared to other whole-genome techniques for phylogenetics, this method seems particularly attractive, as differences in gene content among organisms are readily explicable both in terms of their evolutionary meaning (adaptation to environment) and the mechanisms behind them (gene duplication, gene loss, horizontal gene transfer). In contrast, differences in G-C content, dinucleotide frequencies, and peptide composition—while evidently containing a phylogenetic signal—have no obvious functional or evolutionary interpretation. Since gene content comparisons have more appeal from an evolutionary and functional perspective than other whole-genome methods, and give similar results [57], there is a strong argument for using gene content comparisons as a supplement to sequence-based approaches to phylogenetic analysis.

2.10 Graphs

Graphs are an extremely flexible data structure that allow the modelling of many types of problems. An undirected graph $G = (V, E)$ is composed of a set of vertices, denoted V, which are connected by a set of edges, denoted E. (A set is simply an unordered collection of elements, none of which is repeated). An example of an undirected graph is shown in Figure 2.2. This graph has five vertices labeled a, b, c, d, and e, and has edges between a and b, a and c, a and d, b and c, and c and d. Thus, $V = \{a, b, c, d, e\}$ and $E = \{\{a, b\}, \{a, c\}, \{a, d\}, \{b, c\}, \{c, d\}\}$. Notice that E is a set, and each element of E is itself a set. Since this is an undirected graph, there is no directionality to the edges, and thus each edge is described by a set of vertices. This is appropriate because order is

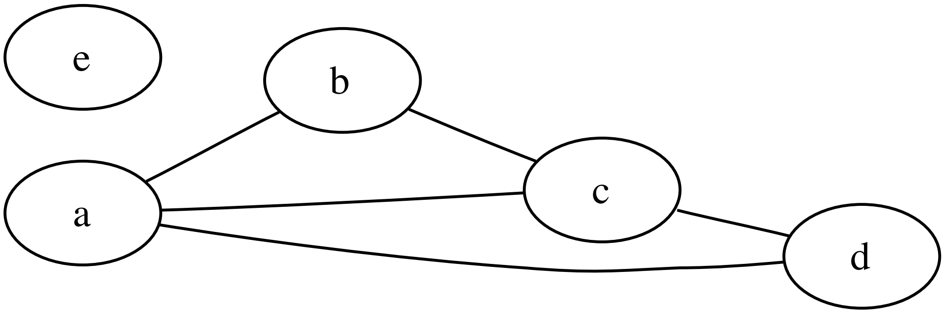

Figure 2.2: Graphical representation of the undirected graph $G = (V, E)$, where $V = \{a, b, c, d, e\}$ and $E = \{\{a, b\}, \{a, c\}, \{a, d\}, \{b, c\}, \{c, d\}\}$.

irrelevant in a set—$\{a, b\}$ is the same as $\{b, a\}$.

Graphs may also be directed. In this case, each edge is specified by an ordered pair rather than a set, wherein the first element of the ordered pair indicates the originating vertex, and the second element denotes the destination vertex. In a visual representation of a directed graph, an arrowhead indicates the destination vertex. Suppose that there is a graph with the same vertices as above ($V = \{a, b, c, d, e\}$), except now $E = \{(a, b), (b, a), (c, d), (d, e), (e, b), (e, c)\}$. A visual representation of this graph is shown in Figure 2.3. Note that Figure 2.3 represents the standard notation for representing directed graphs, which is different than the notation used for visually representing COGs (see Section 2.7.3).

A graph may have one or more connected components. In order to describe a connected component, the concept of a path must first be explained. A path is a sequence of vertices in which each successive vertex is connected to the previous vertex by an edge. A connected component is a set $C \subseteq V$ such that there is a path from every vertex in C to every other vertex in C, and there does not exist a vertex $v \in (V - C)$ such that v is connected by an edge to a vertex in C.

Graphs provide a very natural way of modelling orthologous relationships among proteins. In modelling orthology, each vertex represents a protein, and the presence of an edge between two vertices indicates that the two proteins represented by those vertices are orthologues. Thus, the connected components of such a graph represent groups of orthologous proteins.

An example of this concept is given in Figures 2.4 and 2.5. Suppose that orthology relationships among the proteins in three different organisms have been determined. In Figures 2.4 and 2.5, vertices representing proteins from the same organism are the same color. Since similarity between two proteins from the same organism is not important, there are no edges between like-coloured vertices. In Figure 2.4, it is difficult to identify the connected components using a quick visual inspection, even though the graph is small (only 20 vertices, compared to the thousands of vertices that would be present in a graph representing the proteins from even a small number of bacteria).

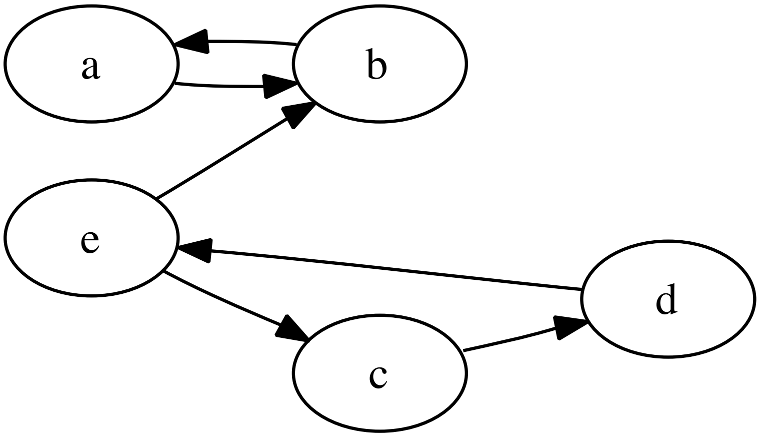

Figure 2.3: Graphical representation of the directed graph $G = (V, E)$, where $V = \{a, b, c, d, e\}$ and $E = \{(a, b), (b, a), (c, d), (d, e), (e, b), (e, c)\}$.

Figure 2.5 shows exactly the same graph, except that the vertices are arranged such that the connected components can be identified easily by visual inspection. This example is designed to give a sense of the fact that finding the connected components of a graph is not as trivial an operation as it may seem (as just looking at Figure 2.5 might suggest), either for a machine or by human visual inspection. The connected components of a graph can be identified using a depth-first search (DFS), which runs in $O(|V| + |E|)$ time.

2.11 Disjoint-set data structures

Depending on the number of vertices and edges, DFS can be quite slow, meaning that the connected components of a large graph could take a long time to compute. Another method for finding orthologous groups once orthologous relationships between pairs of proteins have been determined utilizes a so-called disjoint-set data structure, which is sometimes called a union-find data structure [65]. Disjoint-set data structures can solve a problem equivalent to finding the connected components of a graph because the problem of finding connected components is analogous to the problem of finding disjoint sets (sets that have no elements in common).

A disjoint-set data structure maintains a set of disjoint sets $S = \{S_1, S_2, \ldots, S_n\}$ to which members can be added [65]. Such data structures generally support three operations:

- make–set(x)—creates a new set having x as its only member,
- union(x, y)—forms a new set $S_k = S_i \cup S_j$, where $x \in S_i$ and $y \in S_j$, and destroys the sets S_i and S_j (since all sets must be pairwise disjoint), and
- find–set(x)—find the set containing element x.

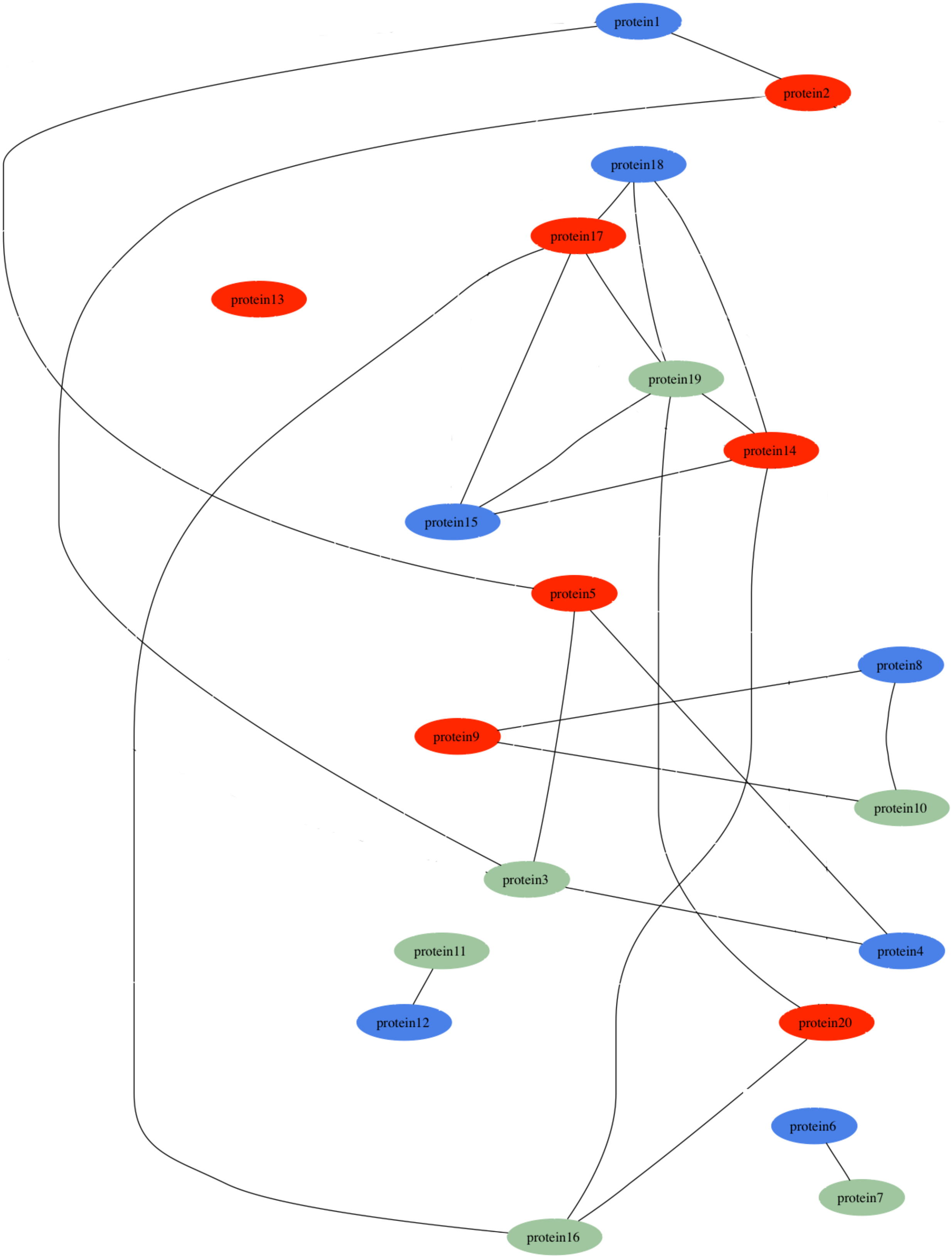

Figure 2.4: An example of a graph for which it is difficult to find the connected components using visual inspection.

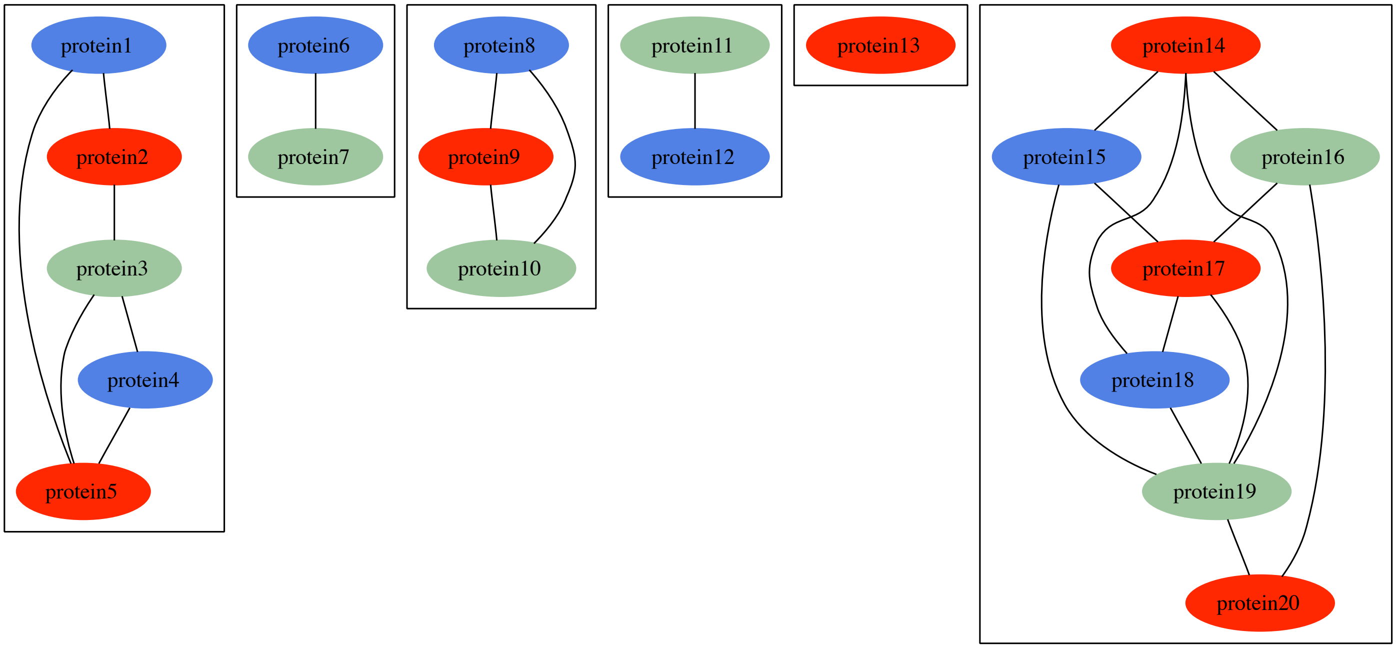

Figure 2.5: The same graph as in Figure 2.4, except the connected components are clearly shown by reorganizing the layout of the graph. Each box contains one connected component.

Once the requisite sequence of union and make–set operations have been performed, finding the orthologous groups is simply a matter of outputting each element of S.

For implementation purposes, each set has a representative element, and all elements of a given set point to the representative element of that set. When two sets are joined using the union operation, the pointer for each element of one of the sets (for efficiency, this is generally the smaller set) are changed to point to the representative element of the other set.

Chapter 3

Research Goal

The wealth of information provided by sequencing efforts, as well as the existence of bioinformatics programs and techniques that facilitate the searching and analysis of large sequence databases, enable organisms to be studied and compared on a genome-wide scale. This thesis, which takes advantage of these sequencing efforts and bioinformatics techniques, has three main goals. The first is to describe the design of a program called PSI, whose purpose is to identify orthologous groups of proteins in sets of organisms, and then to find proteins that are present in some of those organisms, but not others. This goal is described in more detail in Section 3.1. The second goal is to explore the use of PSI for determining what proteins are responsible for a particular phenotype (Section 3.2). Finally, Section 3.3 describes the third goal, which is to investigate the use of PSI for characterizing differences in protein content among different sets of organisms.

3.1 Creating and analyzing the PSI program

This section of the thesis describes the creation of a program called PSI that can, in a very general way, compare the protein content in arbitrary groups of organisms. The method for comparing protein content uses an idea called genome subtraction, which involves "subtracting" the proteins present in one set of organisms from those present in a second set. In other words, PSI identifies proteins that are present in all of the organisms in one set of organisms, but none of the organisms in a second set. For a more formal mathematical description of this operation, see Section 4.1.4.

Among the first to propose the idea of genome subtraction were Huynen et al. [66], who compared the proteins in the pathogen *Haemophilus influenzae* to those in a benign strain of the bacterium *Escherichia Coli* in an attempt to identify genes that might contribute to pathogenicity. It is also possible to perform a "phyletic pattern search" using COGs [32]. However, the dearth of genomes that have been incorporated into COGs, as well as the fact that the percentage of an organism's genome represented in COGs is usually substantially less than 100% (see Section 2.7.3), make COGs unsuitable for performing genome subtraction. Another relevant utility is OrthoMCL [67], which uses RBH to find orthologues, and then uses a Markov clustering algorithm to split apart large groups of orthologues. It also allows the user to find orthologous groups that are present in one set of organisms, but not a second set. Compared to PSI, OrthoMCL implements a more sophisticated orthologue detection technique; on the other hand, it is designed specifically for eukaryotes, rather than prokaryotes, and is difficult to automate easily for large numbers of comparisons.

In addition to describing the design and implementation of PSI, one aspect of its computational efficiency is analyzed. Specifically, the efficiency of DFS on a graph is compared to the use of a disjoint-set data structure for finding orthologous groups of proteins.

3.2 Using PSI to identify protein-phenotype relationships

The biochemical techniques outlined in Section 2.8 for determining what protein(s) may be responsible for a given phenotype are time-consuming and expensive. In addition, a researcher may initially have no idea what proteins are likely candidates for causing the phenotype of interest. Even well-studied organisms contain dozens of uncharacterized proteins, and less well-studied organisms may contain hundreds or even thousands (see also Section 2.8); as such, it is not feasible to create gene knock-in or gene knock-out organisms for every possible protein. Therefore, the ability to narrow down the list of proteins that might be responsible for a given phenotype would be extremely helpful in facilitating the identification of the correct protein or proteins. As stated in Section 3.1, PSI has the ability to identify proteins that are present in one group of organisms (in this case, those that exhibit the phenotype), but not in another group (those that do not exhibit the phenotype). Proteins satisfying both of these criteria are good candidates for causing the phenotype, and the number of these proteins should be substantially smaller than the total number of proteins present in the organisms' proteomes, significantly reducing the number of proteins that must be tested in the laboratory in order to find the protein(s) responsible for the phenotype. In order to test the ability of PSI to find the protein responsible for a given phenotype, a phenotype for which the causative protein is known is selected, and then organisms that exhibit the phenotype and organisms that do not exhibit the phenotype are identified. Then, PSI is used to find proteins that are present in all of the organisms that have the phenotype, and in none of the organisms that do not have the phenotype, to determine whether the correct protein can be identified.

3.3 Using PSI for phylogenetics and comparative genomics

While a large amount of analysis has been done in comparing pairs of organisms (and using this information to construct phylogenetic trees or perform other types of comparative genomics), little work has been done in comparing the genomes of larger groups of organisms. However, a recently developed web server called EDGAR (efficient database framework for comparative genome analysis using BLAST score ratios) [68] allows singlets (proteins occurring in only one organism of a given set) and core proteomes (proteins occurring in all organisms in a given set) to be computed. EDGAR should prove very useful for answering many comparative genomics questions. However, it does have limitations: first, comparisons can only be performed among organisms from the same genus; second, the fact that it is a web server makes it unfeasible to perform hundreds or thousands of comparisons, as some analyses may require. In contrast to EDGAR, PSI is not restricted to intra-genus comparisons, and its stand-alone nature allows it to easily perform hundreds or thousands of comparisons. PSI is also more flexible than EDGAR, as it can find proteins present in any number of organisms that are not present in any number of other organisms, as opposed to only being able to find singlets and core proteomes. (For completeness, it should also be noted that EDGAR can perform functions not possible using PSI, such as the creation of synteny plots, which compare the gene order in several genomes).

PSI has the potential to provide interesting data concerning the similarities and differences in the gene content of individual organisms, of species, of genera, as well as of other arbitrary groups of organisms. Two specific applications of PSI are explored that fall under the categories of phylogenetics and comparative genomics.

First, PSI is used to find the number of unique proteins in pairs of bacteria. This information is then used to calculate a proteomic distance for each pair of bacteria. Based on this information, a phylogenetic tree is created, similar to the procedure performed by Snel et al. [62] (see also Section 2.9.2). This tree encompasses 16 well-studied genera, and represents (to the author's knowledge) the largest phylogenetic tree created using protein content comparisons to date.

Second, PSI is used to study how well-defined bacterial species are from the perspective of protein content. More specifically, two questions are asked: are the isolates of bacterial species X similar to each other in terms of their protein content, and are the isolates of bacterial species X distinct from other isolates of the same genus in terms of their protein content? Answering these questions should provide insight into the quality of existing taxonomic classifications. Assuming that current taxonomic classifications are sound, it is expected that isolates of a given species should be very similar to each other (have many proteins in common), but be quite different from other species (have many proteins not found in other isolates that are from a different species, but the same genus).

Although only these two applications are studied in detail in this thesis, PSI should prove useful for many other comparative genomics applications. Some of these applications are discussed in Section 6.7.

Chapter 4

Data and Methodology

This section describes the methodology used to perform the three main analyses done for this thesis: the creation of the PSI program (Section 4.1), using PSI for the discovery of protein-phenotype relationships (Section 4.2), and using PSI for applications related to phylogenetics and comparing protein content in groups of organisms (Section 4.3).

4.1 Creating and analyzing the PSI program

This section describes the methodology used to create and analyze the PSI program. The algorithm for orthologue detection is presented in Section 4.1.1. Section 4.1.2 describes a technique for visualizing orthologous groups of proteins, while the design of a database containing information necessary for this visualization is given in Section 4.1.3. Section 4.1.4 presents the procedure used to find orthologues that are present in all of the organisms in one set of organisms, but none of the organisms in a second set (the "genome subtraction"). Section 4.1.5 contains the methodology for comparing the computational efficiency of two possible strategies for finding orthologous groups of proteins. Finally, an analytical method for choosing appropriate E-value thresholds for PSI is presented in Section 4.1.6.

4.1.1 Orthologue detection

Choice of orthologue detection method

In order for PSI to be able to identify proteins that are present in all of the organisms in one set of organisms, but none of the organisms in another set, orthologous groups of proteins must first be identified. As described in Section 2.7, many methods for detecting orthologues have been developed. These methods have different strengths and weaknesses, and also have different sensitivities and specificities [47]. Comparing these tools would be quite involved—tree-based tools such as RIO [43], Orthostrapper [44], and OrthologID [42] use different parameters than do BLAST-based methods such as RBH, COGs [23, 33, 34], and INPARANOID [45, 46]. In addition, some of these tools give different types of outputs, making them even more difficult to compare [47].

While it would be of interest to compare the behavior and efficacy of PSI when different orthologue detection methods are used, such an analysis is beyond the scope of this thesis.

Given this, which method should be chosen for orthologue detection? Given their popularity and the extensive manual curation procedures used to create them, COGs would seem like an ideal choice. However, given the limitations described in Section 2.7.3—in particular, the limited number of genomes for which they are available—COGs are not suitable. Some of the other methods that have been described are fairly complex, which could make them complicated to analyze.

The orthologue detection method used for PSI is RBH, which was described in Section 2.7.2. This method has the following advantages.

- RBH is a common and well-understood method that is often used as the basis for more sophisticated approaches to orthologue detection.

- RBH involves only a single tunable parameter (the BLAST E-value threshold), making it straightforward to analyze.

- Although RBH has been found to be less accurate than some other techniques when applied to eukaryotic proteins, it is generally accurate for prokaryotic proteins [47], which are the focus of this thesis.

- Compared to the other simple method of detecting orthologues described in Section 2.7.1, RBH is more accurate, only slightly more complicated, and does not involve any additional parameters.

Performing pairwise BLAST comparisons

Assume that there are n_O organisms labeled $O_1, \ldots, O_{n_O}$. The first step in identifying groups of orthologous proteins is, for each pair of organisms O_i and O_j $(1 \leq i, j \leq n_O; i \neq j)$, and for each protein P_k from organism O_i, to perform a BLAST search using P_k as the query sequence, and all of the proteins in the proteome of O_j as the database. A more formal description of this procedure is shown in Algorithm 4.1. The BioPerl module `Bio::SearchIO` is used to parse the BLAST output.

Creating a graph of orthologous proteins

As described earlier, graphs provide a very convenient way of modelling protein orthology. Each protein in a given organism is represented by a vertex. Two vertices are connected with an edge if their corresponding proteins satisfy the three criteria of RBH (see Section 2.7.2).

The Perl module `Graph.pm` implements the data structure for a graph, and allows one to easily perform simple graph operations, such as adding and deleting vertices and edges, as well as more complex operations, like determining various properties of a graph. The files created in step 7 of Algorithm 4.1 are used to add the appropriate vertices to the graph (one corresponding to each

Algorithm 4.1 Algorithm for performing BLAST queries for all possible pairs of proteins in all possible pairs of organisms. The BioPerl module `Bio::SearchIO` is used to parse the output file from BLAST (step 7).

```
1 Download the proteomes for $O_1, \ldots, O_{n_O}$ in FASTA format
2 Use the formatdb program to create a BLAST database of the proteins in each
   organism's proteome
3 For each organism $O_i$
4     For each protein $P_k$ from $O_i$
5         For each organism $O_j$ $(j \neq i)$
6             Use the blastp program to find the protein $P_l$ that is the best hit when $P_k$ is
              used as a query sequence and the proteins in $O_j$ are used as the database
7             Write the relevant information from the BLAST output file created in step 6
              (query accession number, hit accession number, and E-value) to a tab-delimited file
```

protein), as well as the appropriate edges (when two proteins are predicted orthologues). As the `Graph.pm` module allows attributes of edges to be stored, the E-value of a given comparison is stored as an attribute of the corresponding edge. For each pair of proteins, there are in fact two E-values: one from when the first protein is used as the query sequence, and one from when the second is used. The larger of these two E-values is the one actually stored for a given edge. A more formal description of this procedure is given in Algorithm 4.2.

It should be noted that this implementation of RBH does not attempt to handle ties in the BLAST results. The hits in a BLAST report are sorted in order of ascending E-values; thus, the first hit is the most significant one. In Algorithm 4.1, only the first BLAST hit is saved for a given query protein P_k and database organism O_j; second or subsequent proteins are ignored. However, there could be more than one BLAST result that attains the smallest E-value. Perhaps the most likely situation would be two or more hits having an E-value of 0.0. (It should also be noted that two hits that each receive an E-value of 0.0 could potentially be differentiated on the basis of their bit scores.) Ignoring all but the first BLAST hit for a given query protein could be problematic for orthologue detection. Suppose that the query protein P_k from organism O_i is searched against the database of proteins in the proteome of organism O_j, and that two proteins, denoted P_l^1 and P_l^2, each attain an E-value of 0.0. Also suppose that P_k is the sole best hit in O_i for both P_l^1 and P_l^2. The current implementation of RBH would declare P_k and P_l^1 to be orthologues, but not P_k and P_l^2. However, it would be just as reasonable to call P_k and P_l^2 orthologues. As such, the implementation of RBH described here does not handle this situation correctly; a better strategy would declare both pairs of proteins (P_k and P_l^1, and P_k and P_l^2) as orthologues. However, this modification to the implementation is left as future work.

Once the graph has been created, groups of orthologues are identified by finding the connected

Algorithm 4.2 Algorithm for creating a graph of orthologous proteins.

```
Read the file created in step 7 of Algorithm 4.1 and create a two-dimensional hash of best hits,
  where the first dimension is a query accession, the second dimension is a hit accession, and the
  value is the E-value for the hit between these two proteins
Choose an E-value threshold $T$
For each query protein $P_k$ in the first dimension of the hash
    Create a vertex $V_k$ corresponding to $P_k$ (unless $V_k$ has already been created)
    For each protein $P_l$, the best hit when $P_k$ is used as the query sequence against one of the
      organisms $O_1, \ldots, O_{n_O}$
        Find the corresponding E-value $E_{kl}$
        Create a vertex $V_l$ corresponding to $P_l$ (unless $V_l$ has already been created)
        Find $P_m$, the best hit when $P_l$ is used as the query sequence against the organism
          encoding $P_k$, and the corresponding E-value $E_{lm}$
        If $P_m = P_k$ and $E_{kl} < T$ and $E_{lm} < T$
            Create an edge between $V_k$ and $V_l$ labeled with the value $\max(E_{kl}, E_{lm})$
```

components of the graph (see also Section 2.10). This is done using the `connected_components` method implemented by the `Graph.pm` Perl module.

4.1.2 Visualizing the groups of orthologues

In order to make the results reported by PSI user-friendly and visually pleasing, a visual representation of each group of orthologues is created. This is implemented using the graph visualization tool GraphViz [30], which reads graphs in a text-based format called "dot", and can output a visual representation of that graph in a variety of possible formats (PDF, PNG, and so on). The graphs created by PSI have vertices displaying the following information:

- the Swiss-Prot accession number of the protein,
- the organism that produces the protein,
- the length of the protein,
- a description of the protein,
- a list of keywords describing the protein, and
- a list of gene ontology (GO) terms [69].

A GO term is a standardized term describing a protein, and can fall into one of three categories: "biological process", "molecular function", or "cellular localization". Each vertex lists, for each GO term that applies to that protein, the accession number of the GO term, the description of the GO term, and the category to which the GO term belongs. When PSI is applied to determining protein-phenotype relationships, including the GO terms that describe each protein better enables the user to peruse the candidate proteins and determine which proteins are most likely to be responsible for the phenotype of interest.

In addition, each vertex is a clickable link that takes the user to the UniProt entry for the corresponding protein, where the user can find additional information about the protein that is not present in the vertex itself.

Each edge is labeled with the larger of the two E-values of the two BLAST comparisons (the one in which the first protein is used as the query sequence and the proteome containing the second protein is used as the database, and vice versa). Two vertices representing proteins from the same organism are the same color, whereas vertices representing proteins from different organisms have different colours.

An example of a file in dot format is given in Figure 4.1. This file corresponds to the graph shown in Figure 5.10.

4.1.3 Database of protein information

The input to PSI consists in part of a number of proteomes structured as multi-sequence FASTA files. This file format is chosen because it is the most widely used—and is also the most compact—format for representing sequence information. However, these FASTA files do not contain all of the information that is displayed inside the vertices in the visual representation of the groups of orthologues (see Section 4.1.2); thus, some method must be chosen in order to obtain this information. There are at least three possible solutions to this problem.

1. Accept a file format that does contain all of this information, such as UniProt format. This option suffers from the fact that, depending upon where a particular proteome is downloaded from, a sequence format containing all of the required information may not be available.

2. Use the accession numbers found in the FASTA file to look up sequence information on the internet. For instance, given the accession number Q03BX3, information on this protein can be found by downloading and parsing the HTML from the URL `http://www.uniprot.org/uniprot/Q03BX3`. However, this would necessitate thousands of queries and would therefore be quite slow, even if batch queries were used.

3. Create a relational database containing all protein sequences in Swiss-Prot and TrEMBL, which contain manually annotated and automatically-annotated sequences, respectively [70], and then query this database by a protein's accession number when information about a particular protein is needed.

```
graph graph4437 {
    Q03SD4_387344_taxid [style = filled fillcolor = gold label = "ID: Q03SD4\n
        Organism: Lactobacillus brevis (strain ATCC 367 / JCM 1170)\n
        Length: 342\n
        Description: NADPH:quinone reductase related Zn-dependent oxidoreductase\n
        Keywords: Complete proteome\n
        GO:0008152 (Biological process = metabolic process)\n
        GO:0008270 (Molecular role = zinc ion binding)\n
        GO:0016491 (Molecular role = oxidoreductase activity)\n"
        URL = "http://ca.expasy.org/uniprot/Q03SD4"];
    Q03BX3_321967_taxid [style = filled fillcolor = cyan label = "ID: Q03BX3\n
        Organism: Lactobacillus casei (strain ATCC 334)\n
        Length: 340\n
        Description: NADPH:quinone reductase related Zn-dependent oxidoreductase\n
        Keywords: Complete proteome\n
        GO:0008152 (Biological process = metabolic process)\n
        GO:0008270 (Molecular role = zinc ion binding)\n
        GO:0016491 (Molecular role = oxidoreductase activity)\n"
        URL = "http://ca.expasy.org/uniprot/Q03BX3"];
    Q03DD5_278197_taxid [style = filled fillcolor = darkseagreen label = "ID: Q03DD5\n
        Organism: Pediococcus pentosaceus (strain ATCC 25745 / 183-1w)\n
        Length: 345\n
        Description: NADPH:quinone reductase related Zn-dependent oxidoreductase\n
        Keywords: Complete proteome\nGO:0008152 (Biological process = metabolic process)\n
        GO:0008270 (Molecular role = zinc ion binding)\n
        GO:0016491 (Molecular role = oxidoreductase activity)\n"
        URL = "http://ca.expasy.org/uniprot/Q03DD5"];

    Q03SD4_387344_taxid--Q03BX3_321967_taxid [label = "1e-117" color = gold];
    Q03SD4_387344_taxid--Q03DD5_278197_taxid [label = "1e-125" color = darkseagreen];
    Q03BX3_321967_taxid--Q03DD5_278197_taxid [label = "1e-112" color = darkseagreen];
}
```

Figure 4.1: Example of a file using the dot format for specifying graphs. Each line that specifies an element of the graph ends with a semicolon. There are six such lines in this example; the first three, which denote vertices, are spread out over multiple physical lines. The last three lines specify edges.

The third option appears to be preferable, as accessing these data from a local database would be faster and more reliable than from an external website. Its only disadvantage would be the storage space needed for the database.

For these reasons, a structured query language (SQL) relational database is created using PostgreSQL version 8.3.7. The first step in building this database is to create a database schema. The database schema specifies four tables: “sequences”, “sequences_keywords”, “go_terms”, and “sequences_go_terms”. The “sequences” table contains basic information about a protein, such as its accession number, its length, its description, and so on. The “go_terms” table consists of a list of GO terms, and each entry includes the accession number of that GO term, its type (“molecular role”, “cellular localization”, or “biological process”), and a description of the GO term. Since each sequence can be described by more than one GO term, the table “sequences_go_terms” provides a mapping between sequences and GO terms. Likewise, each sequence can also be described by zero or more keywords, and the table “sequences_keywords” provides a mapping between sequences and keywords. Note that in contrast to GO terms, where each term has associated with it an accession number, a description, and a type, there is no information associated with each keyword other than the keyword itself. As such, there is no need to create a “keywords” table, and no such table is included in the database schema. The schema for this database is given in Figure 4.2.

The second step in creating this database is to download all of the Swiss-Prot and TrEMBL protein sequences. These were downloaded on October 18, 2008 in UniProt format. A Perl script is written to parse these files, extract the relevant information from each record, and output the SQL commands required to enter that information into the database. An example of a Swiss-Prot record, given for the protein with accession number P81928, is given in Figure 4.3. The SQL generated for this protein is shown in Figure 4.4. Part of Figure 4.4 shows the insertion of the GO term with accession number GO:0007275 into the database. If this GO term is encountered again for another protein, it is not inserted into the database again. However, another record would be inserted into the “sequences_go_terms” table. This record would include the accession number of the already-seen GO term, as well as the accession number of the new protein that is also described by this GO term.

The Perl modules DBI (which provides a general interface for interacting with any type of database) and DBD::Pg (which provides routines written specifically to interact with PostgreSQL databases) are used to query the completed database.

4.1.4 Performing the genome subtraction

Mathematical set notation provides a convenient way of describing the concept of genome subtraction as used in this thesis. Let P_1 denote the set of proteins encoded by the genome of organism O_1, P_2 denote the set of proteins encoded by the genome of organism O_2, and so on. The set of all

```
CREATE TABLE sequences (
     accession VARCHAR(40) PRIMARY KEY,
     id VARCHAR(40),
     type VARCHAR(7) NOT NULL CHECK (type IN (‘DNA’, ‘Protein’)),
     length INTEGER NOT NULL,
     molecular_weight NUMERIC,
     sequence VARCHAR(100000) NOT NULL,
     organism_name VARCHAR(2000),
     organism_taxonomy VARCHAR(2000),
     organism_taxid VARCHAR(8),
     description VARCHAR(500) NOT NULL
);

CREATE TABLE sequences_keywords (
     sequence_accession VARCHAR(40) REFERENCES sequences(accession),
     keyword VARCHAR(200),

     PRIMARY KEY (sequence_accession, keyword)
);

CREATE TABLE go_terms (
     accession VARCHAR(20) PRIMARY KEY,
     type VARCHAR(25) NOT NULL CHECK
         (type IN (‘Molecular role’, ‘Cellular localization’, ‘Biological process’)),
     description VARCHAR(200) NOT NULL
);

CREATE TABLE sequences_go_terms (
     sequence_accession VARCHAR(40) REFERENCES sequences(accession),
     go_accession VARCHAR(20) REFERENCES GO_terms(accession),

     PRIMARY KEY (sequence_accession, go_accession)
);
```

Figure 4.2: SQL schema for the database of protein sequences.

```
ID   140U_DROME              Reviewed;         261 AA.
AC   P81928; Q9VFM8;
DT   28-MAR-2003, integrated into UniProtKB/Swiss-Prot.
DT   28-MAR-2003, sequence version 2.
DT   22-JUL-2008, entry version 48.
DE   RecName: Full=RPII140-upstream gene protein;
GN   Name=140up; ORFNames=CG9852;
OS   Drosophila melanogaster (Fruit fly).
OC   Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota;
OC   Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
OC   Ephydroidea; Drosophilidae; Drosophila; Sophophora.
OX   NCBI_TaxID=7227;
CC   -!- FUNCTION: Essential for viability.
CC   -!- SUBCELLULAR LOCATION: Membrane; Multi-pass membrane protein
DR   EMBL; M62975; AAD40352.2; -; Genomic_DNA.
DR   EMBL; AE014297; AAF55023.1; -; Genomic_DNA.
DR   EMBL; AY058577; AAL13806.1; -; mRNA.
DR   PIR; JQ1024; JQ1024.
DR   RefSeq; NP_476951.1; -.
DR   UniGene; Dm.10056; -.
DR   Ensembl; CG9852; Drosophila melanogaster.
DR   GeneID; 41720; -.
DR   KEGG; dme:Dmel_CG9852; -.
DR   NMPDR; fig|7227.3.peg.12715; -.
DR   FlyBase; FBgn0010340; 140up.
DR   HOGENOM; P81928; -.
DR   BioCyc; DMEL-XXX-02:DMEL-XXX-02-011545-MON; -.
DR   ArrayExpress; P81928; -.
DR   GermOnline; CG9852; Drosophila melanogaster.
DR   GO; GO:0007275; P:multicellular organismal development; IMP:UniProtKB.
DR   InterPro; IPR003397; Tim17_Tim22.
DR   Pfam; PF02466; Tim17; 1.
PE   2: Evidence at transcript level;
KW   Complete proteome; Membrane; Transmembrane.
FT   CHAIN         1    261       RPII140-upstream gene protein.
FT                                /FTId=PRO_0000064352.
FT   TRANSMEM     67     87       Potential.
FT   TRANSMEM    131    151       Potential.
FT   TRANSMEM    183    203       Potential.
FT   CONFLICT     64     64       S -> F (in Ref. 1; AAD40352).
SQ   SEQUENCE   261 AA;  29182 MW;  5DB78CF6CFC4435A CRC64;
     MNFLWKGRRF LIAGILPTFE GAADEIVDKE NKTYKAFLAS KPPEETGLER LKQMFTIDEF
     GSISSELNSV YQAGFLGFLI GAIYGGVTQS RVAYMNFMEN NQATAFKSHF DAKKKLQDQF
     TVNFAKGGFK WGWRVGLFTT SYFGIITCMS VYRGKSSIYE YLAAGSITGS LYKVSLGLRG
     MAAGGIIGGF LGGVAGVTSL LLMKASGTSM EEVRYWQYKW RLDRDENIQQ AFKKLTEDEN
     PELFKAHDEK TSEHVSLDTI K
//
```

Figure 4.3: Swiss-Prot entry for the protein with accession number P81928. For brevity, some lines have been removed.

```
INSERT INTO
    sequences
VALUES
    (‘P81928’, ‘140U_DROME’, ‘Protein’, 261, 29182, ‘MNFLWKGRRFLI...’,
    ‘Drosophila melanogaster (Fruit fly)’, ‘Eukaryota; Metazoa; Arthropoda;
    Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota;
    Diptera; Brachycera; Muscomorpha; Ephydroidea; Drosophilidae;
    Drosophila; Sophophora’, ‘7227’,
    ‘RPII140-upstream gene protein’)
;

INSERT INTO
    GO_terms
VALUES
    (‘GO:0007275’, ‘Biological process’, ‘multicellular organismal development’)
;

INSERT INTO
    sequences_GO_terms
VALUES
     (‘P81928’, ‘GO:0007275’)
;

INSERT INTO
    sequences_keywords

VALUES
    (‘P81928’, ‘Complete proteome’)
;

INSERT INTO
    sequences_keywords

VALUES
    (‘P81928’, ‘Membrane’)
;

INSERT INTO
    sequences_keywords

VALUES
    (‘P81928’, ‘Transmembrane’)
;
```

Figure 4.4: SQL statements used to insert the information for the protein with accession number P81928 into the database. For brevity, the amino acid sequence (“MNFLWKGRRFLI...”) has been truncated.

proteins that are found in both O_1 and O_2 can be denoted using set notation as $P_1 \cap P_2$. To clarify this explanation further, suppose that some orthologue detection procedure has been performed, and two orthologous proteins have been identified. For instance, assume that O_1 and O_2 both contain genes coding for cytochrome C oxidase, a protein involved in cellular respiration. Let a represent this protein. Then all of the following are true.

- $a \in P_1$
- $a \in P_2$
- $a \in P_1 \cap P_2$

In other words, if two proteins are orthologous in P_1 and P_2, then they are given the same designation (in this case, a). This does not necessarily mean that the proteins have identical sequences. Note that this is a slight abuse of set notation, as the member, intersection, and union set operations are defined not in terms of strict equality (the same protein from the same organism), but rather in terms of an equivalence relation that deems two proteins to be "equal" if they are orthologues.

As a further example, suppose that protein b is expressed by organism O_1, but does not have an orthologue in the proteome of O_2. Then the following is true.

- $b \in P_1$
- $b \notin P_2$
- $b \notin P_1 \cap P_2$

More generally, the set of proteins that are found in all of the organisms $O = \{O_1, O_2, \ldots, O_{n_O}\}$ (with corresponding proteome sets $P = \{P_1, P_2, \ldots, P_{n_O}\}$) can be denoted as $P_1 \cap P_2 \cap \cdots \cap P_{n_O}$. Suppose that O is partitioned into two disjoint sets $O^I = \{O^I_1, \ldots, O^I_{n_I}\}$ and $O^E = \{O^E_1, \ldots, O^E_{n_E}\}$ such that $O^I \cup O^E = O$ and $O^I \cap O^E = \emptyset$. Therefore, $n_I + n_E = n_O$. Also, let $P^I = \{P^I_1, \ldots, P^I_{n_I}\}$ and $P^E = \{P^E_1, \ldots, P^E_{n_E}\}$ represent the sets of proteomes (with each proteome being a set of proteins) corresponding to the sets of organisms O^I and O^E, respectively. The purpose of PSI is to find the set Z of proteins that are present in all of the proteomes in P^I (the "included" proteomes), but present in none of the proteomes in P^E (the "excluded" proteomes). Proteins in Z will be called "candidate proteins". Then

$$Z = (P^I_1 \cap P^I_2 \cap \cdots \cap P^I_{n_I}) - (P^E_1 \cup P^E_2 \cup \cdots \cup P^E_{n_E}) \tag{4.1}$$

Z could alternatively be expressed as:

$$Z = (P^I_1 \cap P^I_2 \cap \cdots \cap P^I_{n_I}) \cap \sim (P^E_1 \cup P^E_2 \cup \cdots \cup P^E_{n_E}) \tag{4.2}$$

However, the form given in Equation 4.1 gives a more intuitive representation of genome subtraction, since the idea of "subtraction" is suggested by the use of the set difference operator.

In the previous paragraph, the term "candidate protein" was used to refer to a protein that is found in all of the proteomes in P^I, but in none of the proteomes in P^E. It is important to understand that this term refers to a protein with a specific function that may be present in a number of organisms, rather than a single protein from a particular organism. For instance, suppose that one of the candidate proteins (proteins in Z) is cytochrome C oxidase. The term "candidate protein" refers to cytochrome C oxidase as expressed by any of the organisms in O^I, rather than one specific version of this protein from a single organism (say, the one expressed by organism O^I_4). On occasion, the term "candidate group" will be used when there is the need to refer to specific proteins from individual organisms within that group of orthologues. For instance, one could say, "In this candidate group, the protein from organism O^I_4 is 300 amino acids in length." The members of a candidate group are the specific proteins from each organism in O^I that are collectively called a "candidate protein".

Figure 4.5 gives a Venn diagram that illustrates the process of comparing the proteomes of three organisms A, B, and C. PSI can be used to determine which proteins belong to each of the coloured regions shown in this figure. For instance, one might wish to identify:

- proteins that are present only in the proteome of organism A (red area),
- proteins that are present in the proteomes of both organism A and organism B, but not organism C (yellow area), and
- proteins that are present in the proteomes of all three organisms (white area).

In the last case, no proteomes are actually being "subtracted", although the same concepts outlined in this section nonetheless apply, with $O^I = \{A, B, C\}$ and $O^E = \emptyset$. Which regions may be of interest will, of course, depend on the specific biological question that is being considered.

4.1.5 Comparing methods for finding orthologous groups

Once pairwise orthologous relationships have been ascertained using RBH, there are two possible ways to find groups of orthologous proteins. The first is to build a graph and then use DFS to find the connected components (see Section 2.10), and the second is to use a disjoint-set data structure (see Section 2.11). Conveniently, the `graph.pm` Perl module provides a method that finds the connected components of a graph, and also allows the creation of a disjoint-set data structure. Preliminary testing shows that finding orthologous groups by using DFS to determine the connected components of a graph containing many thousands of vertices—as is done in this thesis—takes a significant amount of computational time on an ordinary computer (from seconds

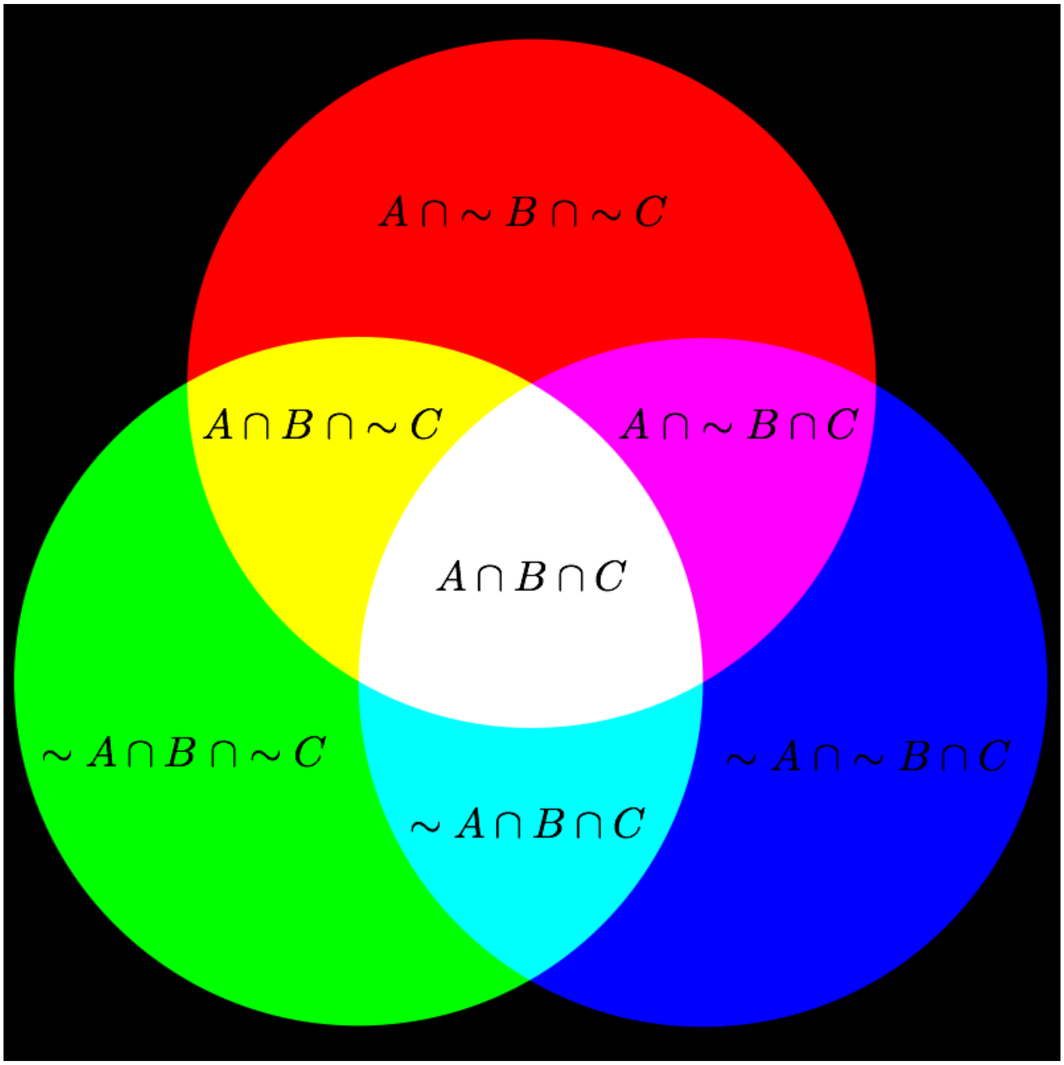

Figure 4.5: Venn diagram representing the sets of proteins found in the proteomes of three organisms (A, B, and C), and the overlap between them. The complete top circle (consisting of the colours red, yellow, pink, and white) represents the set of proteins in the proteome of organism A, while the left circle (green, yellow, light blue, and white) and the right circle (dark blue, light blue, pink, and white) represent the sets of proteins in the proteomes of organisms B and C, respectively. Each individual color is labeled with an expression indicating the set of proteins represented by the region with that color.

or minutes on graphs with a few thousand vertices and edges to hours on graphs with tens of thousands of vertices and edges). This motivates an analysis that determines the fastest way to find groups of orthologues. According to Cormen et al. [65], the use of a disjoint-set data structure is generally faster if the connected components (or disjoint sets) need to be determined repeatedly as elements are added to the data structure, whereas using DFS on a graph is usually preferable if this information only needs to be determined once (after all of the elements have been added to the data structure). However, casual testing suggests that, even when the connected components need to be found only once for a given graph (as in this thesis), the use of the disjoint-set data structure is still much faster.

In order to more rigorously determine the relative speed of building a graph and then using DFS compared to using a disjoint-set data structure, the time taken for each of these two methods is determined using different numbers of organisms. Specifically, tests are performed using the proteins from two isolates of the genus *Streptococcus*, then for three *Streptococcus* isolates, and so on, up to 20 *Streptococcus* isolates. Elapsed times are measured on an otherwise-quiescent Apple iMac with a 2.4 GHz Intel Core 2 Duo processor and four GB of RAM. For each data structure and for each number of *Streptococcus* isolates, measurements are made of both the time taken to build the data structure itself (the graph or the disjoint-set data structure), as well as the time taken to find the connected components using DFS (in the case of the graph) or to find the disjoint sets (in the case of the disjoint-set data structure). Timings were performed in triplicate, and the average of the three times was taken for a given test. The results of these tests are given in Section 5.1.1.

4.1.6 Analytical method for choosing E-value thresholds

The value chosen for the BLAST E-value threshold could have a substantial impact on the correctness of the results produced by PSI. If the threshold chosen is too stringent (small E-values), then two proteins that are actually orthologues may not be identified as such. This could result in both reduced sensitivity and reduced specificity when trying to determine the protein or proteins responsible for a given phenotype. If these proteins are in two organisms that have the phenotype, then they could be missed as possible candidates for causing the phenotype (reduced sensitivity). Conversely, if one protein is found in an organism that has the phenotype and the other is found in an organism that does not, then missing the orthology between these two proteins could result in the protein from the organism having the phenotype being classified as a possible candidate for causing the phenotype, when in fact it should not be (reduced specificity). Choosing an E-value threshold on the opposite side of the spectrum—a very large (non-stringent) E-value threshold—could cause two proteins that are not actually orthologues to erroneously be identified as such, which could also result in both reduced sensitivity and reduced specificity. If the two proteins are from organisms that both exhibit the phenotype, then incorrectly declaring them as orthologues could result in them being erroneously classified as candidate proteins (reduced specificity); conversely, if one of the proteins is from an organism that exhibits the phenotype and one is from an organism that does not, then a possible candidate for causing the phenotype could be missed (reduced sensitivity).

Thus, it is important to choose appropriate E-value thresholds, and there are two possible ways to do this. The first is an analytical method, which considers the number of organisms that are involved in a particular comparison, as well as the number of proteins in the proteome of each organism. The second is an experimental method, using a range of different E-values for a specific comparison to determine the E-values that seem to give good results. The first method is discussed below. The second method is described in Section 4.2.2, where it is analyzed in the context of protein-phenotype relationships, and in Section 4.3.1, where it is analyzed in the context of finding the number of proteins in one organism, but not a second organism.

Suppose that the proteomes of n_o organisms are to be compared. Further suppose that the number of proteins encoded by the organism with the largest proteome in a given comparison is n_p. For each pair of organisms, there will be at most $n_p \times n_p = n_p^2$ pairwise comparisons between proteins. The number of pairs of organisms that must be compared (note that comparisons must be performed in both directions) is $n_o \times (n_o - 1) \approx n_o^2$. Thus, the total number of protein-protein comparisons that must be performed will be bounded above by $n_p^2 n_o^2$. The expected number of spurious matches M will be equal to the number of comparisons performed, multiplied by the probability of a spurious match in each comparison. Let P be the probability of a spurious match. Then

$$M = P n_p^2 n_o^2$$

How can a value for P be derived? The E-value, simply denoted as E in this section, represents for a particular match with raw score S the number of matches attaining a score better than or equal to S that would occur at random given the size of the database. While E does not represent a probability, P can be derived from it: since the probability of finding no random matches with a score greater than or equal to S is e^{-E}, where e is the base of the natural logarithm, the chance of obtaining one or more such matches is $P = 1 - e^{-E}$ [71]. Since P is nearly equal to E when $E < 0.01$, E can reasonably be used as a proxy for P. As such, the expected number of spurious matches M can be written as:

$$M = E n_p^2 n_o^2$$

By rearranging, an equation is obtained that expresses the E-value threshold that should be chosen in terms of n_p, n_o, and M, where M represents the desired value for the expected number of spurious matches:

$$\mathrm{E} = \frac{M}{n_p^2 n_o^2} \tag{4.3}$$

For simplicity, it would be convenient to choose a single E-value threshold that is appropriate for all comparisons done for this thesis. The bacterial species examined in this thesis that has the largest genome, *Burkholderia xenovorans*, encodes $8951 \approx 10^5$ proteins. Thus, a conservative value for n_p would be 10^5, and an upper bound for the greatest number of pairwise comparisons that would take place between two bacteria is $n_p^2 = 10^{10}$. Furthermore, the greatest number of organisms used in a single comparison for this thesis is about 30. Then $n_o = 30$, and the number of pairs of organisms is approximately $n_o^2 = 900$. The total number of pairwise protein comparisons is therefore bounded above by $n_p^2 n_o^2 = 10^{10} \times 900 \approx 10^{13}$. If the expected number of matches that should occur by chance in a given comparison should be one, which is an arbitrarily-chosen but reasonable-sounding choice, then the E-value threshold should be chosen as follows:

$$\mathrm{E} = \frac{1}{10^{10} \times 10^{3}} = 10^{-13}$$

In other words, the matches between two proteins (in both directions) must both have E-values of less than 10^{-13} in order for the proteins to be considered orthologues, in addition to each being the other's best BLAST hit. This E-value threshold is rather conservative, given that most comparisons involve fewer than 30 organisms, and that all of the bacterial proteomes in fact have fewer than 10^5 proteins. Therefore, the actual number of expected spurious matches for all comparisons is, in fact, less than one.

It is interesting to note that the range of E-values that could potentially be chosen using this analytical method is fairly narrow, regardless of the values of each variable in Equation 4.3. The conservative estimate derived above represents one end of the scale, while at the other end of the scale, one might have a comparison involving just two organisms, each having only about 2000 proteins. Assuming $M = 1$, this suggests an E-value threshold of 6.25×10^{-8}. While this represents a difference of over four orders of magnitude, the range of E-values that may be reported by BLAST is far wider, ranging over more than 150 orders of magnitude. This suggests that the actual values of n_o and n_p are not terribly important.

4.2 Using PSI to identify protein-phenotype relationships

This section presents the methodology used to analyze the application of PSI to determining protein-phenotype relationships. Section 4.2.1 describes the use of PSI for determining the protein or proteins responsible for the difference in cell shape between species of the genus *Lactobacillus*, and the related bacterium *Pediococcus pentosaceus*. Section 4.2.2 describes an empirical investigation of the effect of the E-value threshold on the results of the cell shape phenotype comparison, as a complement to the analytical method given in Section 4.1.6. Section 4.2.3 describes the use of PSI for determining proteins that could potentially be involved in the ability of some LAB, but not others, to be resistant to the antibiotic gatifloxacin. Finally, the relationship between the number of organisms having and lacking the phenotype for a given comparison, and the number of candidate proteins, is determined for two cases—one involving the aforementioned cell shape phenotype, and the other involving a hypothetical phenotype present in *Streptococcus* isolates, but not in *Mycobacterium* isolates (Section 4.2.4).

4.2.1 Identifying the protein responsible for cell shape in LAB

MreB is an actin-like protein that is involved in determining bacterial cell shape [72]. Bacteria that contain this protein generally have rod-shaped cells, while those that lack this protein generally have round cells. All of the *Lactobacillus* isolates that have been sequenced to date are rod-shaped, while the related bacterium *P. pentosaceus* is round. See Table A.5 for a list of the 15 lactobacilli (which represent all the lactobacilli that have been sequenced to date) used for this comparison, as well as the number of proteins found in each. The specific *Pediococcus* isolate used is *P. pentosaceus* ATCC 25745, which contains 1755 proteins. PSI is used to identify proteins that are present in all 15 of these lactobacilli, but not in *P. pentosaceus*, to determine whether it is able to discover MreB as a candidate protein. An E-value threshold of 10^{-13} is used, as justified in Section 4.1.6. The results of this analysis can be found in Section 5.2.1.

4.2.2 Evaluating the effect of the E-value threshold on the results of the cell shape phenotype comparison

As a complement to the analytical method presented in Section 4.1.6 for choosing an appropriate E-value threshold, this section describes an empirical analysis of the effect that the E-value threshold has on the results of the cell shape phenotype comparison. In Section 4.2.1, this comparison was performed using just one choice of E-value threshold; in this section, the same comparison is performed using E-value thresholds between 10^0 and 10^{-180}. Data were gathered for each power of ten between these numbers $(10^0, 10^{-1}, 10^{-2}, \ldots, 10^{-199}, 10^{-180})$. A lower limit of 10^{-180} was chosen for the E-value thresholds because BLAST reports E-values smaller than 10^{-180} simply as 0.0. All of the candidate proteins that are reported for any of these E-value thresholds are compiled. For each of these candidate proteins, the range of E-values for which that protein is, in fact, reported as a candidate protein is determined. Results are given in Section 5.2.2.

4.2.3 Identifying the protein(s) conferring gatifloxacin resistance in LAB

Gatifloxacin is an antibacterial drug belonging to the fluoroquinolone family of antibiotics. It works by inhibiting DNA gyrase and DNA topoisomerase IV, which are enzymes involved in facilitating gene expression and DNA replication [73]. Bacterial resistance to this antibiotic is usually associated with point mutations in the genes encoding the target proteins, although other resistance mechanisms, such as efflux pumps, may also play a role. The goal in this section is to identify proteins that might be responsible for gatifloxacin resistance in some LAB. Specifically, *Lactobacillus brevis* and *P. pentosaceus* are known to be resistant to this antibiotic, whereas *Lactobacillus casei* and *Lactobacillus reuteri* F275 are known to be susceptible [Monique Haakensen, personal communication]. PSI is used to identify proteins that are present in both *L. brevis* and *P. pentosaceus*, but absent in both *L. casei* and *L. reuteri*. Unlike the cell shape phenotype discussed in Section 4.2.1, the correct candidate proteins are not yet known. As before, an E-value threshold of 10^{-13} is used. Some candidate proteins for gatifloxacin resistance are shown in Section 5.2.3.

4.2.4 Determining the relationship between the number of input organisms and the number of candidate proteins

For the sake of efficiently finding the proteins responsible for the phenotype of interest, it is important that PSI be able to produce a reasonably short list of candidate proteins. Recall that the set of organisms having the phenotype and the set of organisms lacking the phenotype were denoted by O^I and O^E, respectively, in Section 4.1.4. The values of $|O^I|$ and $|O^E|$—the number of organisms having and lacking the phenotype, respectively—are presumably important in determining the number of candidate proteins that PSI returns. To investigate the relationship between the values of $|O^I|$ and $|O^E|$ and the number of candidate proteins returned, two separate tests are performed.

Test 1—Cell shape phenotype

The first comparison uses as its basis the cell shape phenotype comparison described in Section 4.2.1, which seeks to find the protein responsible for causing the cell shape of *P. pentosaceus* to be different than that of 15 *Lactobacillus* isolates. To do this, proteins are found that are present in all 15 of the lactobacilli, but not in *P. pentosaceus*; thus, $|O^I| = 15$ and $|O^E| = 1$. To determine the effect of $|O^I|$ on the number of candidate proteins returned, $|O^I|$ is varied between one and 15, while $|O^E|$ is kept constant (O^E always contains just *P. pentosaceus*). In other words, for the first comparison ($|O^I| = 1$), PSI is used to determine the proteins that are present in one of the *Lactobacillus* isolates, but not in *P. pentosaceus*, and the number of candidate proteins is recorded. For the second comparison ($|O^I| = 2$), PSI is used to determine the proteins that are present in both proteomes of two *Lactobacillus* isolates, but not in *P. pentosaceus*, and the number of candidate proteins is again recorded. This pattern continues until $|O^I| = 15$. The *Lactobacillus* isolates are added in the same order as they are listed in Table A.5; thus, the first comparison involves only *L. acidophilus*; the second comparison involves both *L. acidophilus* and *L. brevis*, and so on. Note that somewhat different results would be obtained if the lactobacilli were added in a different order. Determining how the order in which the organisms are added affects the results is beyond the scope of this thesis, but would make worthwhile future work.

Test 2: Proteins in *Streptococcus* isolates that are not in *Mycobacterium* isolates

The previous investigation only examines how the number of candidate proteins varies with $|O^I|$, but it is also desirable to determine how this quantity varies with both $|O^I|$ and $|O^E|$. Suppose that the goal is to find the protein responsible for a hypothetical phenotype possessed by all isolates of the genus *Streptococcus*, and none of the isolates of the genus *Mycobacterium.* The genera *Streptococcus* and *Mycobacterium* were selected because each has many isolates sequenced from a variety of species. There are 31 sequenced *Streptococcus* isolates and 14 sequenced *Mycobacterium* isolates; thus, using all isolates, $|O^I| = 31$ and $|O^E| = 14$. However, in this section, the number of candidate proteins is determined for each possible combination of $|O^I|$ and $|O^E|$, where $|O^I|, |O^E| \geq 1$. The isolates are added in a manner analogous to that used in test 1. The results of both test 1 and test 2 are given in Section 5.2.4.

4.3 Using PSI for phylogenetics and comparative genomics

The second main application of PSI explored in this thesis involves analyzing phylogenetic relationships based on protein content, as well as comparing commonalities and differences in protein content in different groups and classifications of bacteria. Section 4.3.1 discusses an empirical investigation into choosing appropriate E-value thresholds when finding the number of proteins found in one organism, but not a second organism. The bacteria used for the comparisons performed in Sections 4.3.3 and 4.3.4 are described in Section 4.3.2. Section 4.3.3 examines the application of PSI to creating phylogenetic trees on the basis of protein content. Finally, PSI is used to determine whether different isolates from the same species are soundly clustered based on their protein content—in other words, it is determined whether the isolates of a given genus are, in terms of protein content, similar to each other, as well as distinct from other isolates of the same genus. The methodology for this analysis is given in Section 4.3.4.

4.3.1 Evaluating the effect of the E-value threshold on numbers of unique proteins

To get a sense of the impact that the choice of E-value threshold has on how many proteins are reported to be in organism A but not organism B (or vice versa), pairs of organisms A and B are selected, and the number of proteins in the proteome of organism A but not in organism B is determined using the same range of E-value thresholds as in Section 4.2.2 ($10^0, 10^{-1}, 10^{-2}, \ldots, 10^{-199}$, 10^{-180}). As always, reciprocal best BLAST hits are required for two proteins to be declared orthologues. Necessarily, the greater the E-value threshold, the fewer unique proteins will be reported. At an E-value threshold of 10^0, for instance, many proteins will (sometimes spuriously) be identified

Table 4.1: Comparisons performed for determining the effect of the E-value threshold on the number of proteins found in one proteome, but not a second proteome.

#	**Proteins found in the proteome of...**	**...that are not found in the proteome of...**
	Intra-species comparisons	
1a	*Pseudomonas putida* GB-1	*Pseudomonas putida* KT2440
1b	*Xanthomonas campestris* 8004	*Xanthomonas campestris* B100
1c	*Staphylococcus aureus* COL	*Staphylococcus aureus* JH1
	Inter-species comparisons	
2a	*Burkholderia mallei* ATCC 23344	*Burkholderia xenovorans* LB400
2b	*Vibrio cholerae* ATCC 39315	*Vibrio fischeri* ATCC 700601
2c	*Streptococcus thermophilus* ATCC BAA-250	*Streptococcus pyogenes* MGAS2096
	Inter-genus comparisons	
3a	*Bacillus anthracis* Ames ancestor	*Corynebacterium diphtheriae* ATCC 700971
3b	*Mycobacterium tuberculosis* ATCC 25177	*Neisseria meningitidis* 053442
3c	*Yersinia enterocolitica* 8081	*Clostridium tetani* E88

as orthologues, resulting in relatively few unique proteins; conversely, at a threshold of 10^{-180}, few proteins will be identified as orthologues (even if they really are orthologues), resulting in many proteins being reported as unique.

It is reasonable to expect that the relatedness of the organisms involved in a comparison would affect the interaction between the E-value threshold and numbers of unique proteins reported. As such, three different degrees of relatedness are considered—two isolates from the same species; two isolates from the same genus but different species; and two isolates from different genera. These are referred to, respectively, as intra-species, inter-species, and inter-genus comparisons. Three pairs of organisms are selected for each of these three types of comparisons. The specific isolates used for each comparison are selected arbitrarily. The comparisons performed are found in Table 4.1, and the results are given in Section 5.3.1.

4.3.2 Bacteria used

The bacteria used for Sections 4.3.3 and 4.3.4 are found in Table 4.2. These genera were selected on the basis of having two or more species that each had two or more isolates sequenced. All of the bacterial proteomes were downloaded from Integr8 [18] on November 28, 2008. A more detailed list of the bacteria used, including information on the genome size and the number of proteins in each isolate, can be found in Appendix A.

Table 4.2: Summary of bacteria used for the analyses described in Sections 4.3.3 and 4.3.4. A detailed listing of the isolates from each genus can be found in the tables comprising Appendix A.

Genus	**Isolates (#)**	**Species (#)**
Bacillus	16	10
Brucella	8	5
Burkholderia	19	10
Clostridium	19	10
Lactobacillus	15	12
Mycobacterium	14	11
Neisseria	6	2
Pseudomonas	15	7
Rhizobium	6	4
Rickettsia	11	9
Shigella	7	4
Staphylococcus	18	4
Streptococcus	31	9
Vibrio	8	5
Xanthomonas	8	3
Yersinia	12	3

4.3.3 Phylogenetics based on protein content

As described in Section 2.9.1, phylogenetic analyses are typically done by comparing 16S rRNA gene sequences or by MLSA. However, there has been increasing interest in using whole-genome approaches to study phylogenetics, as then more information is taken into account than just a single gene or a small number of genes. This section describes a variant of an analysis done by Snel et al. [62] for using protein content to infer evolutionary relationships. As a measure of the proteomic similarity between two organisms, Snel et al. [62] used the number of shared proteins between two organisms divided by the number of proteins in the smaller proteome. In this section, a different proteomic distance metric is proposed that fits within the framework of PSI, as obtaining data for this metric involves genome subtraction. This metric is calculated as the average of the number of proteins in bacterium A that are not in bacterium B, and the number of proteins in bacterium B that are not in bacterium A. This will be called the average unique proteins (AUP) metric.

Using an E-value threshold of 10^{-13}, as justified in Section 4.1.6, the number of proteins in bacterium A but not bacterium B is determined for all pairs of bacteria in Table 4.2. The AUP metric is then calculated for each pair. The unweighted pair group method with arithmetic mean (UPGMA) is used to create a phylogenetic tree. These results are presented in Section 5.3.2.

4.3.4 Evaluating taxonomic classifications by determining how well species are clustered based on protein content

The purpose of this section is to analyze the quality of current taxonomic classifications from a novel perspective—specifically, by determining the level of cohesiveness in the protein content of a given species. Evaluating the taxonomic classifications of different species by examining their protein content could be conceptualized as a clustering problem. The general idea behind clustering is that each element in a given cluster should be similar to other elements in the same cluster, but dissimilar to elements from other clusters. In the context of taxonomy and protein content, the clustering of a given species could be considered sound if two criteria are satisfied: first, the organisms of the species are similar to each other (i.e., have a large core proteome); second, they are distinct from other organisms (i.e., have many proteins found only in that species). To determine whether existing taxonomic classifications fit these criteria, PSI will be used to answer the following questions.

- Is the number of proteins in the core proteome of a particular species having N_I sequenced isolates larger than the core proteome of N_I randomly selected organisms from the same genus?

- Is the number of proteins that are found in all N_I isolates of a given species, but none of the other organisms from the same genus, larger than the number of proteins found in N_I randomly selected isolates of that genus, but no others?

Algorithm 4.3 Algorithm for determining whether a given species has a larger core proteome size than randomly selected sets of isolates from the same genus.

```
1  From genus $G$, choose a species $S$ having $N_I$ sequenced isolates, where $N_I \geq 2$
2  Determine the number of proteins $C_S$ in the core proteome of these $N_I$ isolates
3  $R \leftarrow \emptyset$
4  Do until $|R| = 25$
5      Randomly choose, without replacement, $N_I$ isolates from genus $G$ to form a set $X$
6      If all of the isolates in $X$ are from the same species, go back to step 5
7      If $X$ is equal to any of the sets in $R$, go back to step 5
8      $R \leftarrow R \cup \{X\}$
9      Determine the size of the core proteome of the isolates in $X$
10 Find the average core proteome size $C_R$ of all the sets in $R$, as well as their standard deviation.
11 Perform a t-test to determine whether $C_R$ is significantly different from $C_S$.
```

The rationale behind asking these question is as follows. One would expect that the isolates of a given species would have more proteins in their core proteome, and more unique proteins, than randomly selected sets of isolates from the same genus. Thus, a “yes” answer to each of the above questions would lend support to the species’ current taxonomic classification. In contrast, “no” answers would suggest that the species does not fit the clustering criteria given above, and its taxonomic classification may therefore warrant reexamination.

These questions are answered for each of the species from the genera listed in Table 4.2 that have two or more isolates sequenced. As the methodology used to approach the two questions is somewhat complex, it is presented in two different ways. The following paragraph contains a description of the methodology, and Algorithm 4.3 conveys the same methodology using an algorithmic format. Both of these descriptions apply only to answering the first question; however, the methodology used to answer the second question is analogous, and is briefly described in the final paragraph of this section.

Once again, let N_I be the number of isolates that have been sequenced for a particular species. First, a set of N_I isolates from that genus is randomly selected. This set is examined to ensure that all of its members are not from the same species (either the species being examined, or any other species from that genus). For instance, when generating random sets of two organisms each corresponding to the two *Bacillus thuringiensis* isolates ($N_I = 2$), there should not be a random set containing both *B. thuringiensis* isolates, nor should there be a random set containing two *Bacillus anthracis* isolates. However, a random set containing one *B. thuringiensis* isolate and one *B. anthracis* would be valid. If a random set is generated, but all of its members are from the same species, then the set is discarded and another is generated in its place. PSI is then used to find the size of the core proteome of this set of organisms. This procedure is then repeated 25 times;

in other words, 25 random sets of N_I organisms are constructed, and the size of the core proteome is determined for each. The 25 sets are also checked to ensure that none of the sets are the same. The reasons for choosing 25 random sets, rather than some other quantity, were:

- this number is large enough that the results will be statistically meaningful,
- the computational time to generate the results for 25 random sets was reasonable, and
- this number is not too much larger than the maximum number of random sets that could be generated for some species (see below).

Some genera have too few sequenced isolates to enable 25 sets to be created. For instance, the genus *Neisseria* has only six isolates sequenced in total, with two *Neisseria gonorrhoeae* isolates and four *Neisseria meningitidis* isolates. With respect to generating random sets corresponding to *N. gonorrhoeae*, the number of possible ways to choose two items from six is $\mathrm{C}(6, 2) = 15$. However, seven of these sets have both organisms from the same species, leaving just eight valid sets. Thus, for *N. gonorrhoeae*, line 4 in Algorithm 4.3 would be changed to "Do until $|R| = 8$". Similarly, in generating random sets corresponding to *N. meningitidis*, the number of ways in which one can choose four items from six is the same: $\mathrm{C}(6, 4) = 15$. One of these sets (the one containing all four *N. meningitidis* isolates) is invalid, leaving 14 sets. Besides these two *Neisseria* species, other species for which fewer than 25 sets could be constructed are *Brucella suis* (24 sets), *Rhizobium leguminosarum* (13 sets), *Rhizobium etli* (13 sets), and *Shigella boydii* (17 sets).

After finding the core proteome sizes of all 25 (or fewer for the aforementioned species) random sets for a given species, a t-test is performed to determine whether the mean of the core proteome sizes for the randomly-generated sets is different than the core proteome size of the N_I isolates of the species in question.

The approach to the second question is analogous to the procedure given above, except that rather than finding proteins that are found in all members of a given set of organisms, proteins are found that exist in all members of a given set, *and* in no other organisms from the same genus. The results of all analyses described in this section are give in Section 5.3.3.

Chapter 5

Results

This section of the thesis presents results concerning the creation of PSI (Section 5.1), the use of PSI for determining protein-phenotype relationships (Section 5.2), and the use of PSI for phylogenetics and protein content comparisons (Section 5.3). Note that most of the figures in this section use scalable vector graphics, so the reader may zoom in to see the full detail of each figure if this thesis is being read electronically.

5.1 Creating and analyzing the PSI program

This section consists of a single subsection, 5.1.1, which describes the results of comparing the speed of two different methods for finding orthologous groups.

5.1.1 Comparing methods for finding orthologous groups

This section describes the results of tests determining the relative efficiency of two different methods for finding orthologous groups once the orthologous relationships between pairs of proteins have been determined (see Section 4.1.5 for full methodology). The first method was to build a graph, and then find the connected components of the graph; the second method was to build a disjoint-set data structure, and then find the disjoint sets. It was first verified that both methods gave the same results. Two aspects of their speed were then examined: the time taken to build the data structure, and the time taken to actually find the connected components (for the graph method) or the disjoint sets (for the disjoint-set data structure method). The results of these comparisons are found in Figure 5.1, which compares the time taken to build each data structure when the proteins from between two and 20 *Streptococcus* isolates were used, and Figure 5.2, which compares the time needed to find the orthologous groups once the data structures have been created.

Figure 5.1 shows that the time taken to build the two data structures were similar, with the disjoint-set data structure taking slightly longer to create than the graph when there were many *Streptococcus* isolates. However, the substantive difference can be seen in Figure 5.2, which shows that the disjoint-set data structure was much faster at finding the orthologous groups (by finding the disjoint sets) than the graph data structure (by finding the connected components). In fact, the

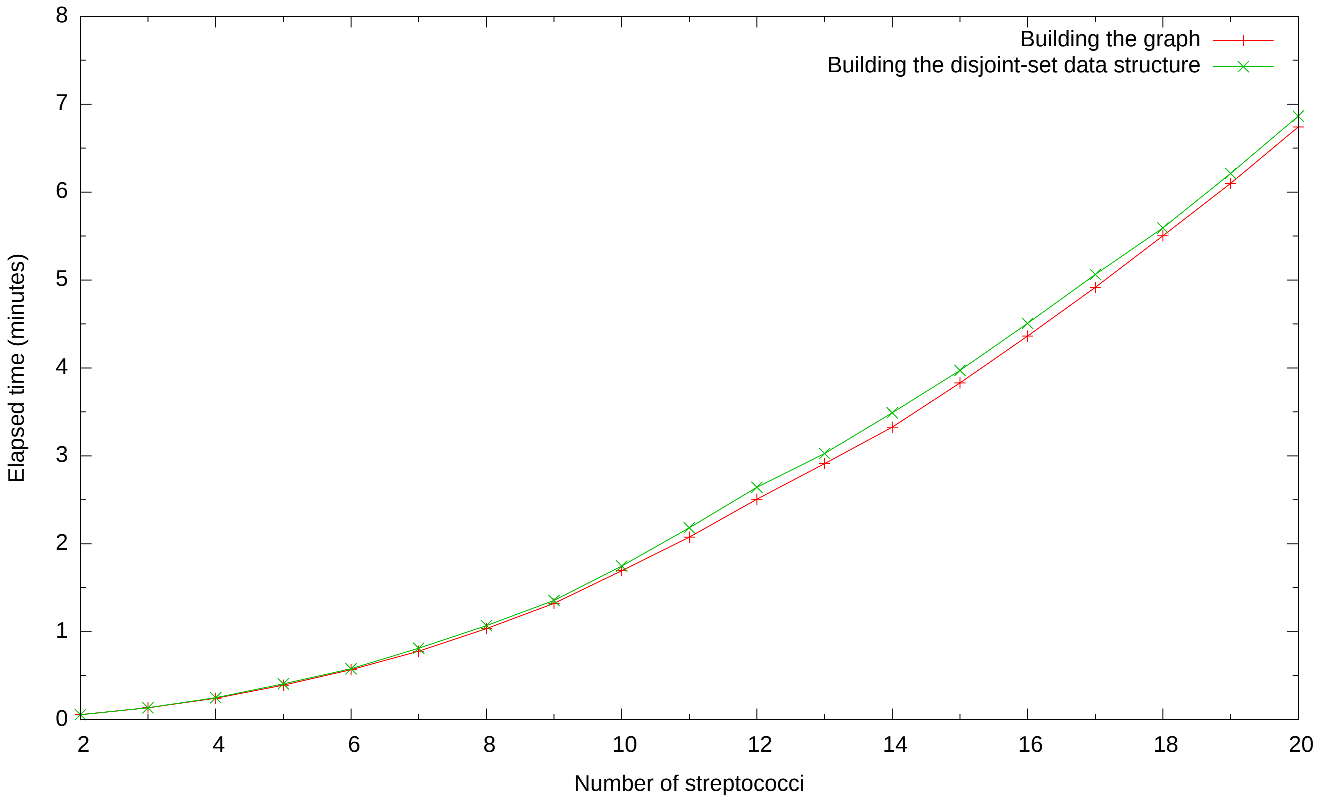

Figure 5.1: Time taken to build the graph data structure or the disjoint-set data structure when finding orthologous groups for between two and 20 *Streptococcus* isolates.

time needed for the disjoint-set data structure to find the actual disjoint sets once the data structure had been created was rather trivial; when 20 *Streptococcus* isolates were used, this operation took just three seconds (note that the plot for the disjoint-set data structure in Figure 5.2 is largely indistinguishable from the x-axis). This was in stark contrast to the time needed to find the connected components of the graph, which took over nine hours. Thus, a disjoint-set data structure proved to be a far better choice than a graph, as it took only slightly longer to build, and was able to output the orthologous groups far more quickly.

The fact that a disjoint-set data structure works well for PSI is likely due to the fact that (to use graph terminology) an application of PSI always has many connected components (groups of orthologous proteins) and few vertices in each connected component (each orthologous group contains a fairly small number of proteins). For instance, when 20 *Streptococcus* isolates were used, there were 39241 vertices (proteins) and 6645 connected components (orthologous groups), for an average of just 5.9 vertices per connected component. This means that the number of union operations that need to be performed when building the disjoint-set data structure, as well as the time taken to perform each union operation, would be relatively small. (See Section 2.11 for background on the union operation, as well as the other operations involved in building a disjoint-set data structure.) To reinforce this idea, suppose that another problem also involves 39241 vertices, but ends up with just 100 connected components after all the union operations have been done. In this case, many union operations would need to be performed, and each union operation would be

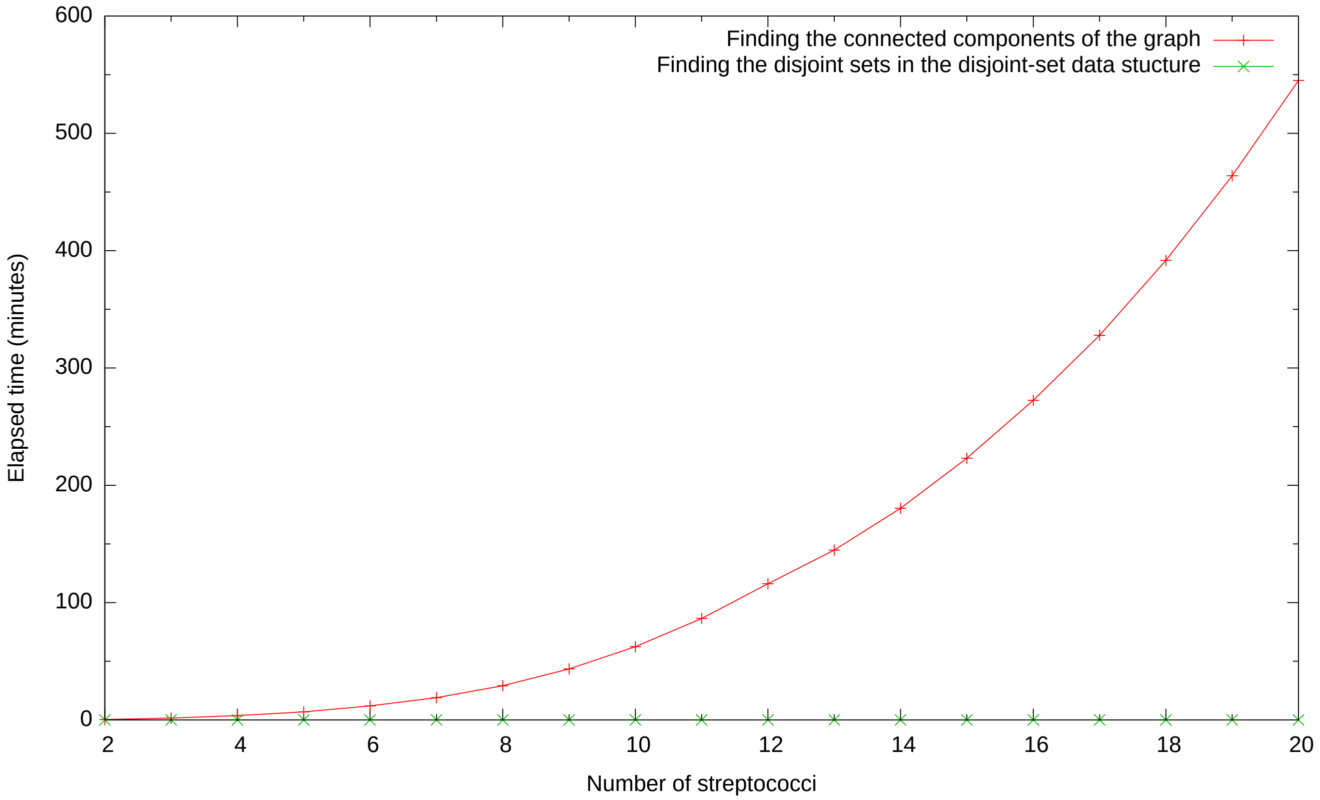

Figure 5.2: Time taken to find the connected components of the graph, or to find the disjoint sets in the disjoint-set data structure, when finding orthologous groups for between two and 20 *Streptococcus* isolates. Note that the plot for the disjoint-set data structure is difficult to distinguish from the x-axis.

fairly time consuming, since the existing disjoint sets would be very large. As the union operation is the most computationally expensive procedure when building a disjoint-set data structure, the fact that an application of PSI will have many connected components and therefore require few union operations makes a disjoint-set data structure ideally suited for it.

While it is easy to understand why the disjoint-set data structure performs well, it is more puzzling why DFS on a graph performs so poorly. DFS should take $O(|V|+|E|)$ time, which is the same as the time complexity needed to build the graph itself. Thus, it is difficult to understand why building the graph took just minutes for 20 *Streptococcus* isolates, but finding the connected components took hours. The code for the `graph.pm` module may provide clues, but unfortunately it contains no internal documentation, and also makes use of other modules for performing the traversal, making it difficult to determine why finding the connected components is so inefficient. Determining the cause of this inefficiency would make worthwhile future work.

The time complexity of a disjoint-set data structure is difficult to express directly in terms of $|V|$ and $|E|$; however, it can be expressed in terms of the number of make-set, union, and find-set operations. Using a linked-list implementation of a disjoint-set data structure, a sequence of m make-set, union, and/or find-set operations, where $|V|$ of these are make-set operations, has a time complexity of $O(m + |V| \log |V|)$ [65]. Most importantly, in PSI the number of find-set operations will be exactly $|V|$, which compares favourably to the $O(|V| + |E|) \approx O(|V|^2)$ time taken to find the connected components of the graph.

5.2 Using PSI to identify protein-phenotype relationships

This section describes the results of using PSI to identify protein-phenotype relationships. In Section 5.2.1, the results of using PSI for finding the protein or proteins responsible for the difference in cell shape between *Lactobacillus* isolates and *P. pentosaceus* are presented. An analysis of how the choice of E-value threshold affects the results of the cell shape phenotype comparison is given in Section 5.2.2. Section 5.2.3 identifies candidate proteins that may be responsible for resistance to the antibiotic gatifloxacin in some LAB. Finally, Section 5.2.4 examines the relationship between the number of input organisms and the number of candidate proteins for two different test cases.

5.2.1 Identifying the protein responsible for cell shape in LAB

As described in Section 4.2.1, PSI was used to find proteins present in all 15 *Lactobacillus* isolates, but not in *P. pentosaceus*. The total number of proteins in all organisms was 32456, while the number of orthologous protein groups created by the orthologue detection procedure described in Section 4.1.1 was 8792. The number of candidate groups—groups containing proteins from all of the *Lactobacillus* isolates, but none from *P. pentosaceus*—was just nine. With the exception of one, each of these candidate groups contained exactly one protein from each *Lactobacillus* isolate. The remaining candidate group, which contained uncharacterized proteins, had two proteins from both strains of *L. casei*, and thus contained a total of 17 proteins rather than 15.

Table 5.1 shows the accession number and description of the protein from *Lactobacillus gasseri* for each of these candidate groups. Proteins from *L. gasseri* were chosen as representatives for each candidate group because, for some candidate groups, the description of the protein from this species was more informative than the descriptions of the proteins from other species. For instance, for one candidate group, the description of the protein from *L. gasseri* was "Lon-like protease with PDZ domain", whereas the description of the protein from *L. acidophilus* from the same candidate group was "Putative uncharacterized protein". Subsequently, a given candidate group will be referred to by the accession number of the *L. gasseri* protein from that group.

There are two proteins in Table 5.1 that appear most likely to be responsible for the difference in cell shape—Q042L8 ("Actin-like ATPase for cell morphogenesis") and Q042N8 ("Rod shape-determining protein MreB"). The descriptions of these two proteins both appear to be very good matches for the phenotype being examined. It may be tempting to conclude that proteins in the Q042N8 candidate group may be more likely to be responsible for the rod shape of *Lactobacillus* species than proteins in the Q042L8 candidate group, since "Rod shape-determining protein MreB"

Table 5.1: Accession number and description of the protein from *Lactobacillus gasseri* for each of the nine candidate groups that result when finding proteins present in all 15 *Lactobacillus* isolates, but not *Pediococcus pentosaceus.*

Accession number	**Description**	**Proteins in candidate group (#)**
Q040U1	Effector of nucleoid occlusion Noc	15
Q042L8	Actin-like ATPase for cell morphogenesis	15
Q046W3	Predicted secreted protein	15
Q042M6	Putative uncharacterized protein	17
Q043A5	Orotate phosphoribosyltransferase	15
Q042Q3	Cell division protein sepF	15
Q045N9	Predicted hydrocarbon binding protein	15
Q042S7	Lon-like protease with PDZ domain	15
Q042N8	Rod shape-determining protein MreB	15

explicitly mentions a rod shape, whereas "Actin-like ATPase for cell morphogenesis" is less specific. However, examining the descriptions of other proteins in both candidate groups dispels this notion. In the Q042N8 candidate group, the protein from *L. brevis* (for instance) is described as "Actin-like ATPase for cell morphogenesis", and in the Q042L8 candidate group, the protein from *Lactobacillus plantarum* (for instance) is described as "Cell shape determining protein MreB". It therefore seems reasonable to suggest that proteins in both groups may be involved in determining cell shape. It should also be noted that, while the presence of the MreB (or MreB-like) proteins are the likeliest explanation for the rod-like shape of these *Lactobacillus* isolates, it is also possible that the cell division protein sepF plays a role, since cell shape is determined during cell division. Furthermore, some of the other proteins listed in Table 5.1 could also play a role in determining cell shape. However, there are likely other phenotypic differences between the lactobacilli and *P. pentosaceus* besides cell shape, and some of the proteins in Table 5.1 could be responsible for these differences.

Of particular note is the fact that this procedure was able to identify the MreB (or MreB-like) proteins regardless of their annotation. Some of the proteins in the two MreB-containing groups contained very specific annotations, such as "Rod shape-determining protein MreB". For other proteins, the annotation mentioned cell shape, but did not specifically call the protein MreB ("Cell shape determining protein"). Furthermore, the degree to which PSI was able to narrow down the list of possible candidate proteins was quite impressive: of 8792 groups of orthologues, only nine contained proteins from all of the *Lactobacillus* isolates, but not *P. pentosaceus.* Clearly, searching through nine candidate proteins for the one most likely to cause the phenotype of interest is much better than having to search through hundreds or thousands of proteins.

In solving this problem, it was assumed that the presence of a protein (or proteins) induces a rod-like cell shape, whereas the absence of this protein results in a round cell shape. Because of this assumption, PSI was used to find proteins that are present in all of the lactobacilli, but not in *P. pentosaceus*. Had the opposite assumption been made—that the presence of a protein induces a round cell shape, while its absence results in a rod-like cell shape—then PSI would have been used to look for proteins found in *P. pentosaceus*, but in none of the lactobacilli. This would likely have been unsuccessful, although laboratory experiments would be necessary to confirm that none of the proteins found in *P. pentosaceus* (but not the lactobacilli) are involved in determining its round cell shape. Biochemical knowledge, or even just intuition, can therefore be very helpful in choosing (to use the notation of Section 4.1.4) which organisms should be in O^I and which should be in O^E. In the absence of *a priori* knowledge or intuition, the best strategy would be to try both possibilities.

The full graph for the candidate group containing Q042N8 is shown in Figure 5.3, while a larger view of a portion of this graph is given in Figure 5.4. Figure 5.4 illustrates the features of the visualizations produced by PSI, as described in Section 4.1.2. Note that in Figure 5.3, the vertices representing proteins from the two *L. reuteri* isolates do not contain the information that the other vertices do. The two *L. reuteri* proteomes were downloaded from IMG [38–40], rather than EBI, and the IMG proteomes do not contain Swiss-Prot/TrEMBL accession numbers. These proteomes were downloaded from IMG rather than EBI because the strain designations at EBI were ambiguous, and an application of PSI not described in this thesis required the precise strains to be identified for this species.

5.2.2 Evaluating the effect of the E-value threshold on the results of the cell shape phenotype comparison

Using the methodology described in Section 4.2.2, it was determined how the E-value threshold affects the results of the cell shape phenotype comparison. The first question was: how does the number of candidate proteins change with the choice of E-value threshold? A plot showing this relationship is given in Figure 5.5 for E-value thresholds ranging from 10^0 to 10^{-180}. The plot shows that the number of candidate proteins decreased fairly steadily as the E-value threshold was decreased, from between nine and twelve at high E-value thresholds (10^0 to 10^{-27}) to between one and two at low E-value thresholds (10^{-134} to 10^{-171}). No candidate proteins were reported when E-value thresholds of less than 10^{-171} were used.

It may initially seem strange that, as the E-value threshold was decreased, there were local oscillations in the number of candidate groups. For instance, there were four candidate groups when an E-value threshold of 10^{-120} was used, three candidate groups when E-value thresholds between 10^{-121} and 10^{-123} were used, and four candidate groups when a threshold of 10^{-124} was

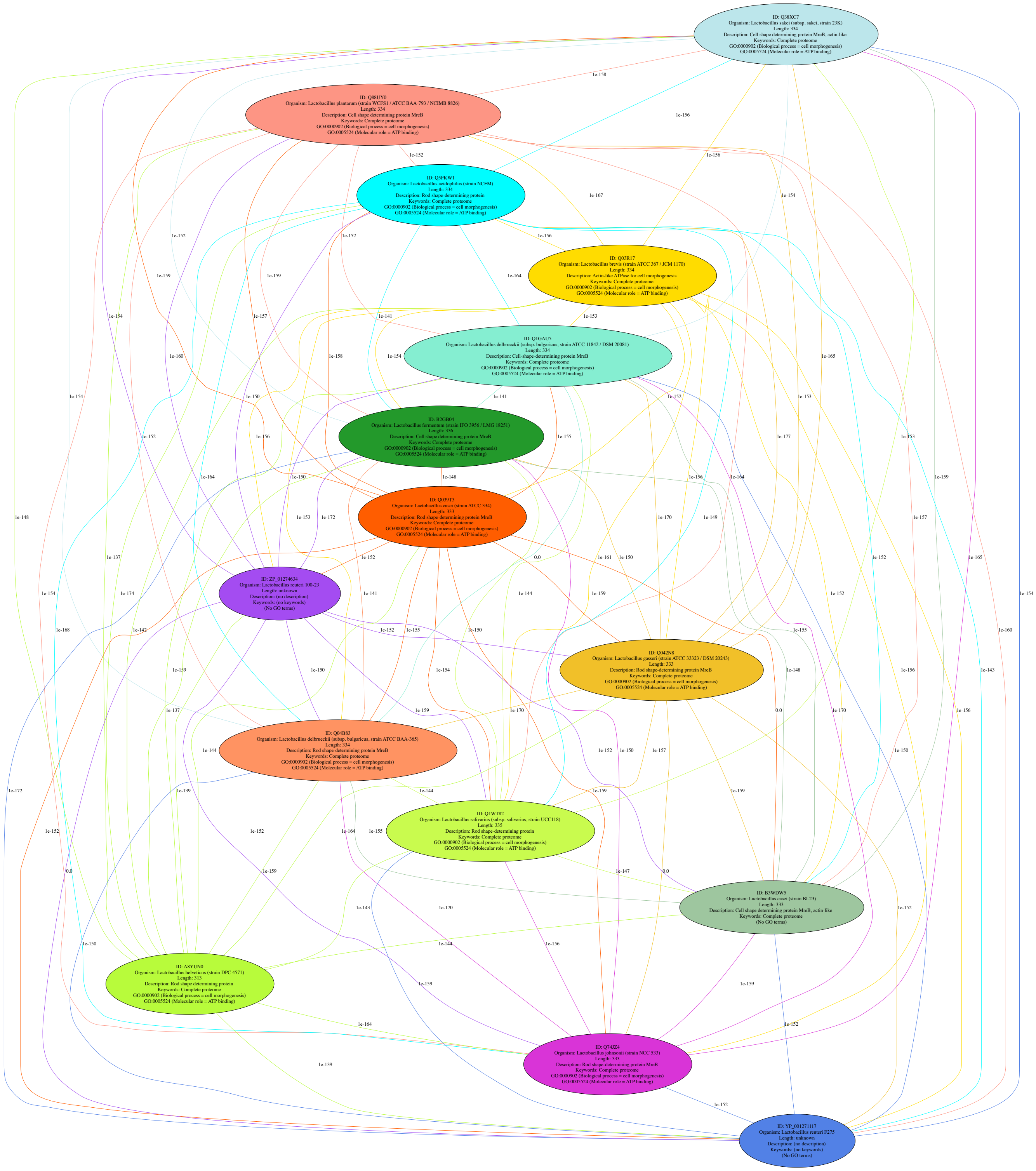
ID: Q38XC7
Organism: Lactobacillus sakei (subsp. sakei, strain 23K)
Length: 334
Description: Cell shape determining protein MreB, actin-like
Keywords: Complete proteome
GO:0000902 (Biological process = cell morphogenesis)
GO:0005524 (Molecular role = ATP binding)
ID: Q88UY0
Organism: Lactobacillus plantarum (strain WCFS1 / ATCC BAA-793 / NCIMB 8826)
Length: 334
Description: Cell shape determining protein MreB
Keywords: Complete proteome
GO:0000902 (Biological process = cell morphogenesis)
GO:0005524 (Molecular role = ATP binding)
ID: Q5FKW1
Organism: Lactobacillus acidophilus (strain NCFM)
Length: 334
Description: Rod shape-determining protein
Keywords: Complete proteome
GO:0000902 (Biological process = cell morphogenesis)
GO:0005524 (Molecular role = ATP binding)
ID: Q03R17
Organism: Lactobacillus brevis (strain ATCC 367 / JCM 1170)
Length: 334
Description: Actin-like ATPase for cell morphogenesis
Keywords: Complete proteome
GO:0000902 (Biological process = cell morphogenesis)
GO:0005524 (Molecular role = ATP binding)
ID: Q1GAU5
Organism: Lactobacillus delbrueckii (subsp. bulgaricus, strain ATCC 11842 / DSM 20081)
Length: 334
Description: Cell-shape-determining protein MreB
Keywords: Complete proteome
GO:0000902 (Biological process = cell morphogenesis)
GO:0005524 (Molecular role = ATP binding)
ID: B2GB04
Organism: Lactobacillus fermentum (strain IFO 3956 / LMG 18251)
Length: 336
Description: Cell shape determining protein MreB
Keywords: Complete proteome
GO:0000902 (Biological process = cell morphogenesis)
GO:0005524 (Molecular role = ATP binding)
ID: Q039T3
Organism: Lactobacillus casei (strain ATCC 334)
Length: 333
Description: Rod shape-determining protein MreB
Keywords: Complete proteome
GO:0000902 (Biological process = cell morphogenesis)
GO:0005524 (Molecular role = ATP binding)
ID: ZP_01274634
Organism: Lactobacillus reuteri 100-23
Length: unknown
Description: (no description)
Keywords: (no keywords)
(No GO terms)
ID: Q042N8
Organism: Lactobacillus gasseri (strain ATCC 33323 / DSM 20243)
Length: 333
Description: Rod shape-determining protein MreB
Keywords: Complete proteome
GO:0000902 (Biological process = cell morphogenesis)
GO:0005524 (Molecular role = ATP binding)
ID: Q04B83
Organism: Lactobacillus delbrueckii (subsp. bulgaricus, strain ATCC BAA-365)
Length: 334
Description: Rod shape-determining protein MreB
Keywords: Complete proteome
GO:0000902 (Biological process = cell morphogenesis)
GO:0005524 (Molecular role = ATP binding)
ID: Q1WT82
Organism: Lactobacillus salivarius (subsp. salivarius, strain UCC118)
Length: 335
Description: Rod shape-determining protein
Keywords: Complete proteome
GO:0000902 (Biological process = cell morphogenesis)
GO:0005524 (Molecular role = ATP binding)
ID: B3WDW5
Organism: Lactobacillus casei (strain BL23)
Length: 333
Description: Cell shape determining protein MreB, actin-like
Keywords: Complete proteome
(No GO terms)
ID: A8YUN0
Organism: Lactobacillus helveticus (strain DPC 4571)
Length: 313
Description: Rod shape determining protein
Keywords: Complete proteome
GO:0000902 (Biological process = cell morphogenesis)
GO:0005524 (Molecular role = ATP binding)
ID: Q74IZ4
Organism: Lactobacillus johnsonii (strain NCC 533)
Length: 333
Description: Rod shape-determining protein MreB
Keywords: Complete proteome
GO:0000902 (Biological process = cell morphogenesis)
GO:0005524 (Molecular role = ATP binding)
ID: YP_001271117
Organism: Lactobacillus reuteri F275
Length: unknown
Description: (no description)
Keywords: (no keywords)
(No GO terms)
1e-158
1e-156
1e-152
1e-156
1e-154
1e-167
1e-152
1e-152
1e-156
1e-159
1e-159
1e-164
1e-154
1e-157
1e-141
1e-153
1e-165
1e-160
1e-158
1e-154
1e-150
1e-154
1e-152
1e-141
1e-153
1e-156
1e-155
1e-152
1e-177
1e-153
1e-148
1e-150
1e-164
1e-156
1e-164
1e-159
1e-148
1e-153
1e-172
1e-170
1e-149
1e-157
1e-137
0.0
1e-161
1e-150
1e-152
1e-165
1e-152
1e-154
1e-174
1e-141
1e-144
1e-159
1e-160
1e-168
1e-142
1e-152
1e-155
1e-150
1e-155
1e-159
1e-148
1e-156
1e-143
1e-150
1e-154
1e-137
1e-170
1e-159
0.0
1e-170
1e-156
1e-144
1e-152
1e-150
1e-157
1e-150
1e-172
1e-139
1e-144
1e-159
1e-152
1e-164
1e-155
0.0
1e-147
0.0
1e-152
1e-143
1e-170
1e-150
1e-144
1e-156
1e-159
1e-159
1e-152
1e-164
1e-139
1e-152

Figure 5.3: Complete graph for one of the candidate groups containing the cell shape-determining protein MreB. Each vertex represents a protein, and edges between vertices denote orthologous relationships between proteins. If this thesis is being read electronically, the reader may zoom in on the figure so the details can be viewed.

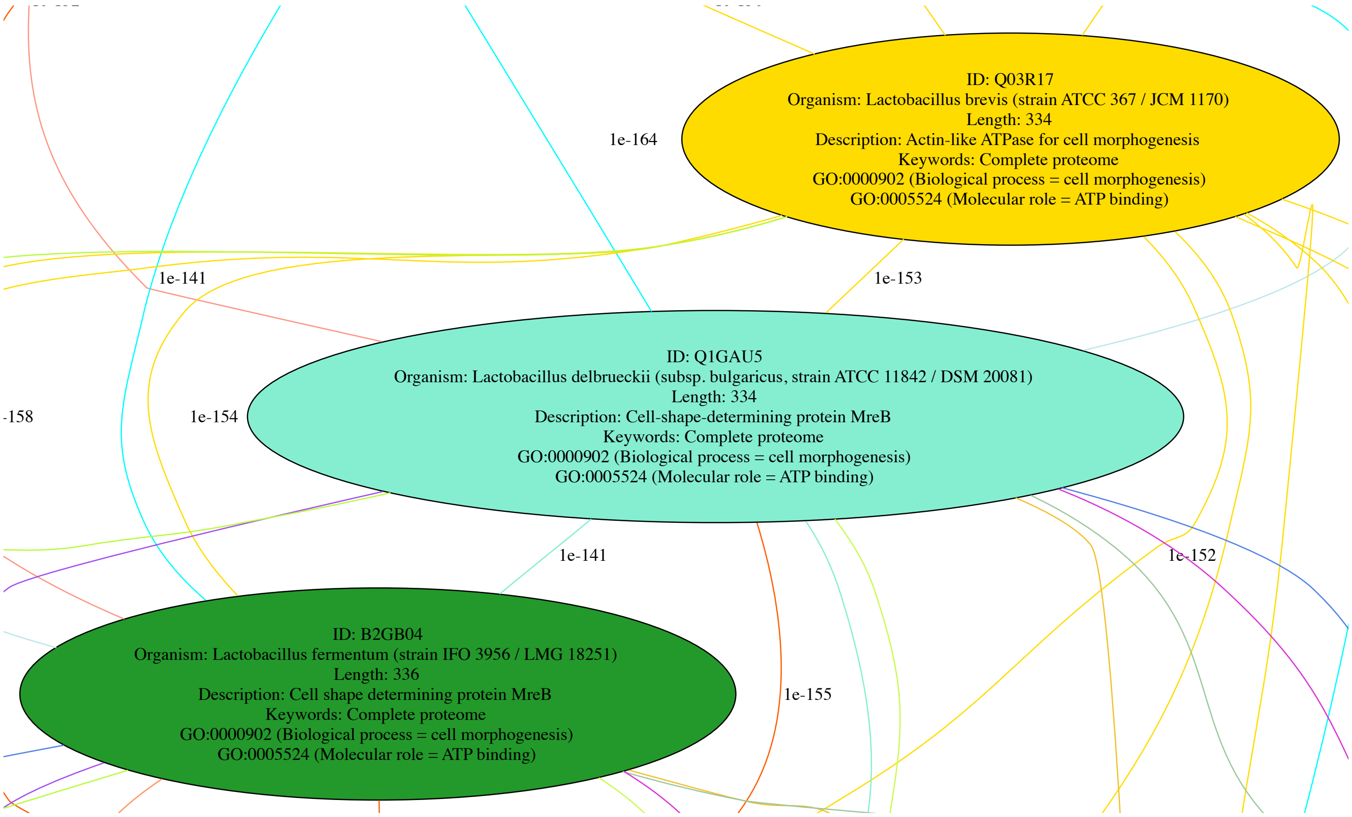
ID: Q03R17
Organism: Lactobacillus brevis (strain ATCC 367 / JCM 1170)
Length: 334
Description: Actin-like ATPase for cell morphogenesis
Keywords: Complete proteome
GO:0000902 (Biological process = cell morphogenesis)
GO:0005524 (Molecular role = ATP binding)
1e-164
1e-141
1e-153
ID: Q1GAU5
Organism: Lactobacillus delbrueckii (subsp. bulgaricus, strain ATCC 11842 / DSM 20081)
Length: 334
Description: Cell-shape-determining protein MreB
Keywords: Complete proteome
GO:0000902 (Biological process = cell morphogenesis)
GO:0005524 (Molecular role = ATP binding)
-158
1e-154
1e-141
1e-152
ID: B2GB04
Organism: Lactobacillus fermentum (strain IFO 3956 / LMG 18251)
Length: 336
Description: Cell shape determining protein MreB
Keywords: Complete proteome
GO:0000902 (Biological process = cell morphogenesis)
GO:0005524 (Molecular role = ATP binding)
1e-155

Figure 5.4: A smaller section of the graph shown in Figure 5.3.

used. Such oscillations can be attributed to vagaries in the BLAST E-values between proteins in pairs of lactobacilli, or between a protein from a *Lactobacillus* isolate and one from *P. pentosaceus*. It is important to understand that decreasing the E-value threshold can either increase or decrease the number of candidate groups. An orthologous group that was previously a candidate group would no longer be a candidate group if a decrease in the E-value threshold causes a protein from a particular *Lactobacillus* isolate to no longer be in that group. Conversely, an orthologous group that was not previously a candidate group may become one if a decrease in the E-value threshold causes a protein from *P. pentosaceus* to no longer be in that group.

All of the candidate proteins that were reported for at least one choice of E-value threshold were then compiled. There were 45 such proteins. For each candidate group x, the range of E-value thresholds for which x was reported as a candidate group was determined; these data are depicted in Figure 5.6. In Figure 5.6, each candidate group is identified using the accession number of the protein from *Lactobacillus helveticus* in that group. The choice of organism was arbitrary; the candidate groups could have been labeled with the accession numbers from any of the other lactobacilli. Table 5.2 gives the description of the protein corresponding to each accession number.

Figure 5.6 shows that many of the 45 candidate groups were reported as such for only a very small range of E-value thresholds. In fact, 29 of the 45 candidate groups were reported for a range of E-value thresholds spanning less than ten orders of magnitude. It should be noted again that, although a range of, say, 10^{-35} to 10^{-40} actually constitutes a seemingly large range of six orders of

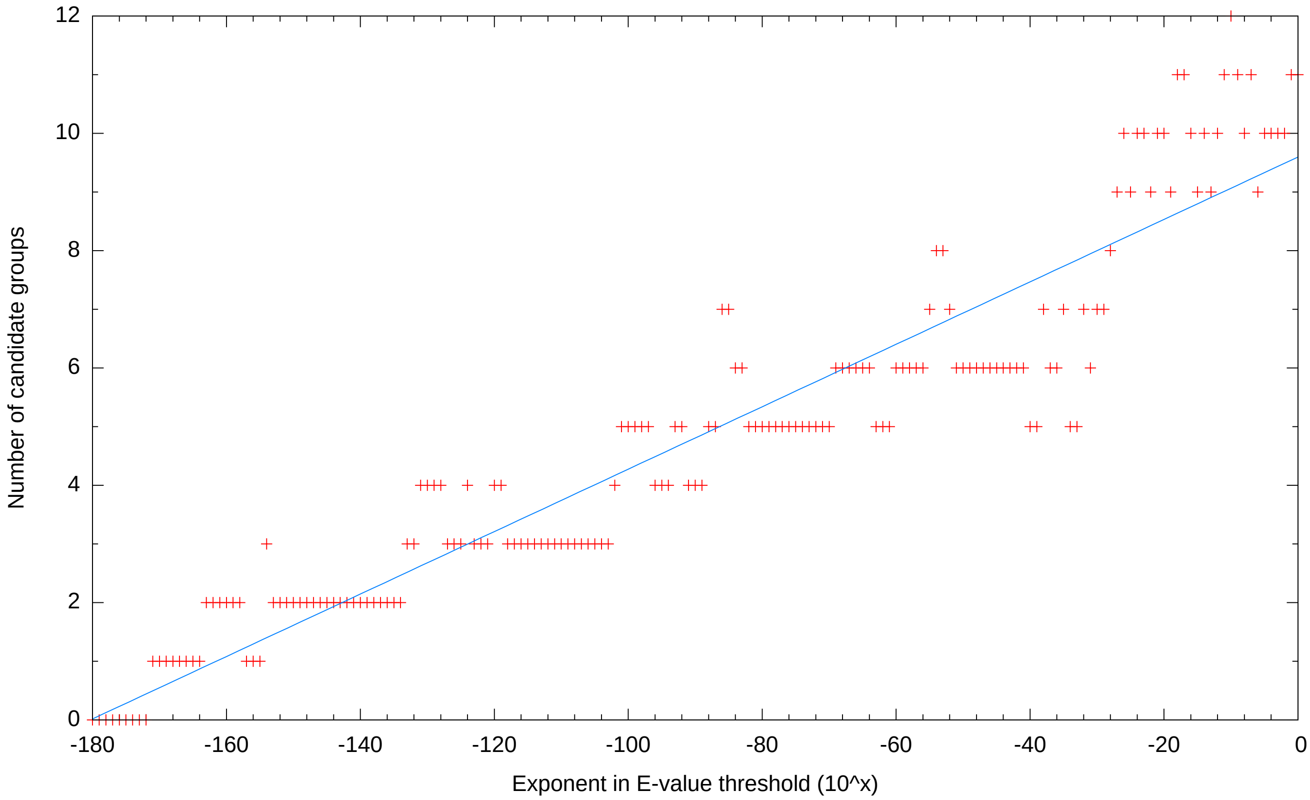

Figure 5.5: Relationship between the E-value threshold and the number of candidate groups for the cell shape phenotype comparison. The best-fit line was calculated with gnuplot using least squares.

magnitude, it actually represents a narrow range relative to the possible E-values that BLAST can report. Given that these 29 candidate groups were reported over such a narrow range of E-value thresholds, it begs the question as to whether these candidate groups should be considered spurious. An analysis of some of these candidate groups suggests that they should.

For instance, consider the candidate group corresponding to the *L. helveticus* protein with accession number A8YWH1, which was only reported as a candidate group for E-value thresholds ranging from 10^{-35} to 10^{-38}. In this range, the group containing A8YWH1 had one protein from each *Lactobacillus* isolate, but no proteins from *P. pentosaceus* (this is true by definition, because it was a candidate group). When an E-value threshold of 10^{-34} was used, the group containing A8YWH1 (which was no longer a candidate group) contained all of the same proteins as when thresholds between 10^{-35} and 10^{-38} were used, but also contained a protein from *P. pentosaceus*. At an E-value threshold of 10^{-39}, the A8YWH1 group contained proteins from just six *Lactobacillus* isolates. The graph containing A8YWH1 produced when the threshold was set at 10^{-34} illustrates why A8YWH1 was part of a candidate group when the E-value threshold was between 10^{-35} and 10^{-38}, but not otherwise. This graph is given in Figure 5.7. The second vertex from the left represents a protein from *P. pentosaceus*. This vertex had only one incident edge, with an E-value of 4×10^{-35}, which is greater than 10^{-35} but less than 10^{-34}, thus explaining the absence of this group as a candidate group when the E-value threshold was 10^{-34}. As stated earlier, the group containing A8YWH1 had only six total proteins when the E-value threshold was set to 10^{-39}. The

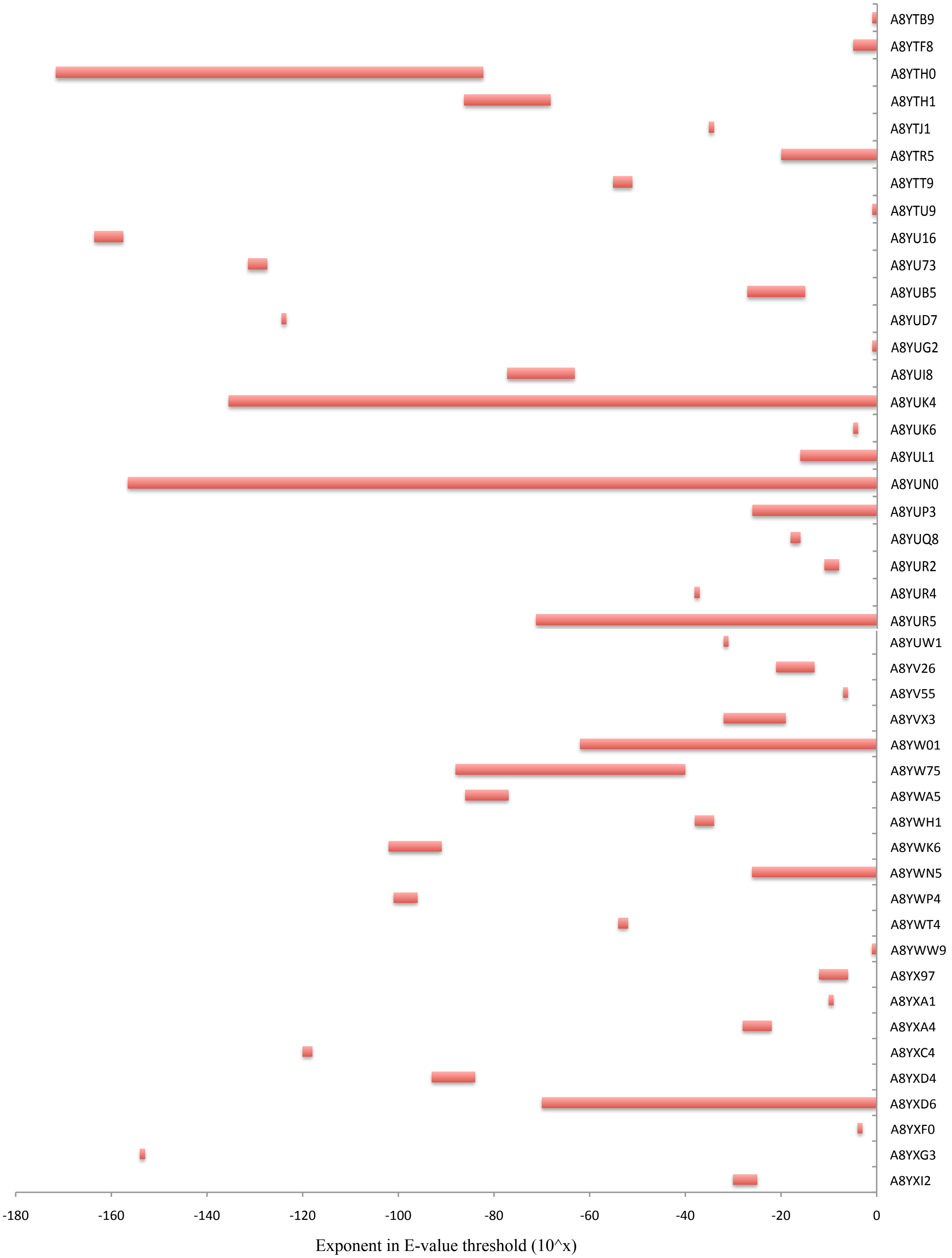

Figure 5.6: Plot representing the range of E-value thresholds for which each candidate group (reported for at least one choice of E-value threshold between 10^0 and 10^{-180}) was, in fact, reported as a candidate group for the cell shape phenotype comparison.

Table 5.2: Description corresponding to the accession number of the protein from *Lactobacillus helveticus* from each of the 45 candidate groups reported using any E-value threshold between 10^0 and 10^{-180} for the cell shape phenotype comparison.

Accession	Description
A8YTB9	Cation efflux protein
A8YTF8	Putative uncharacterized protein
A8YTH0	ABC transporter, ATP-binding protein
A8YTH1	Putative DNA binding protein
A8YTJ1	UPF0297 protein lhv_0439
A8YTR5	Putative uncharacterized protein
A8YTT9	SsrA-binding protein
A8YTU9	Putative uncharacterized protein
A8YU16	Uracil permease
A8YU73	UTP-glucose-1-phosphate uridylyltransferase
A8YUB5	Putative transcriptional regulator
A8YUD7	Putative uncharacterized protein
A8YUG2	Protein-tyrosine phosphatase
A8YUI8	Putative esterase
A8YUK4	Cell shape determining protein
A8YUK6	Putative uncharacterized protein
A8YUL1	Putative uncharacterized protein
A8YUN0	Rod shape determining protein
A8YUP3	Cell division protein sepF
A8YUQ8	UPF0356 protein lhv_0877
A8YUR2	Putative uncharacterized protein
A8YUR4	Phosphopantetheine adenylyltransferase
A8YUR5	Putative uncharacterized protein
A8YUW1	Dihydrofolate reductase
A8YV26	Putative reductase
A8YV55	Protein crcB homolog
A8YVX3	Putative alkaline shock protein
A8YW01	Orotate phosphoribosyltransferase
A8YW75	Putative permease
A8YWA5	tRNA delta(2)-isopentenylpyrophosphate transferase
A8YWH1	Putative ABC transporter
A8YWK6	Putative response regulator
A8YWN5	Putative extracellular protein
A8YWP4	Glutamine ABC transporter ATP binding protein
A8YWT4	Putative sugar kinase
A8YWW9	Putative uncharacterized protein
A8YX97	Putative uncharacterized protein
A8YXA1	Putative uncharacterized protein
A8YXA4	Putative membrane protein
A8YXC4	Sensor protein
A8YXD4	Chromosome partitioning protein
A8YXD6	Chromosome partitioning protein
A8YXF0	Hydrolase of alpha-beta family
A8YXG3	Xanthine permease
A8YXI2	Holo-[acyl-carrier-protein] synthase

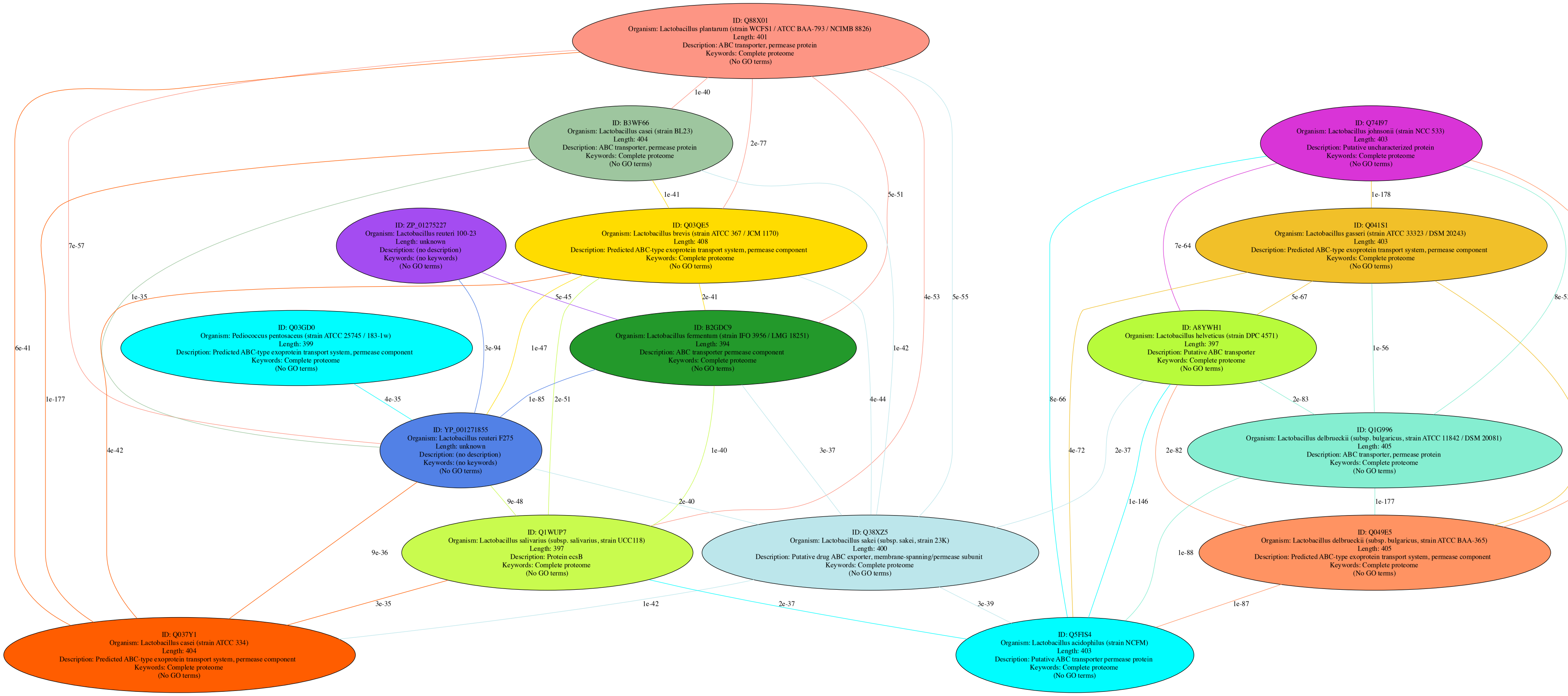
ID: Q88X01
Organism: Lactobacillus plantarum (strain WCFS1 / ATCC BAA-793 / NCIMB 8826)
Length: 401
Description: ABC transporter, permease protein
Keywords: Complete proteome
(No GO terms)
ID: B3WF66
Organism: Lactobacillus casei (strain BL23)
Length: 404
Description: ABC transporter, permease protein
Keywords: Complete proteome
(No GO terms)
ID: Q74I97
Organism: Lactobacillus johnsonii (strain NCC 533)
Length: 403
Description: Putative uncharacterized protein
Keywords: Complete proteome
(No GO terms)
ID: ZP_01275227
Organism: Lactobacillus reuteri 100-23
Length: unknown
Description: (no description)
Keywords: (no keywords)
(No GO terms)
ID: Q03QE5
Organism: Lactobacillus brevis (strain ATCC 367 / JCM 1170)
Length: 408
Description: Predicted ABC-type exoprotein transport system, permease component
Keywords: Complete proteome
(No GO terms)
ID: Q041S1
Organism: Lactobacillus gasseri (strain ATCC 33323 / DSM 20243)
Length: 403
Description: Predicted ABC-type exoprotein transport system, permease component
Keywords: Complete proteome
(No GO terms)
ID: Q03GD0
Organism: Pediococcus pentosaceus (strain ATCC 25745 / 183-1w)
Length: 399
Description: Predicted ABC-type exoprotein transport system, permease component
Keywords: Complete proteome
(No GO terms)
ID: B2GDC9
Organism: Lactobacillus fermentum (strain IFO 3956 / LMG 18251)
Length: 394
Description: ABC transporter permease component
Keywords: Complete proteome
(No GO terms)
ID: A8YWH1
Organism: Lactobacillus helveticus (strain DPC 4571)
Length: 397
Description: Putative ABC transporter
Keywords: Complete proteome
(No GO terms)
ID: YP_001271855
Organism: Lactobacillus reuteri F275
Length: unknown
Description: (no description)
Keywords: (no keywords)
(No GO terms)
ID: Q1G996
Organism: Lactobacillus delbrueckii (subsp. bulgaricus, strain ATCC 11842 / DSM 20081)
Length: 405
Description: ABC transporter, permease protein
Keywords: Complete proteome
(No GO terms)
ID: Q1WUP7
Organism: Lactobacillus salivarius (subsp. salivarius, strain UCC118)
Length: 397
Description: Protein ecsB
Keywords: Complete proteome
(No GO terms)
ID: Q38XZ5
Organism: Lactobacillus sakei (subsp. sakei, strain 23K)
Length: 400
Description: Putative drug ABC exporter, membrane-spanning/permease subunit
Keywords: Complete proteome
(No GO terms)
ID: Q049E5
Organism: Lactobacillus delbrueckii (subsp. bulgaricus, strain ATCC BAA-365)
Length: 405
Description: Predicted ABC-type exoprotein transport system, permease component
Keywords: Complete proteome
(No GO terms)
ID: Q037Y1
Organism: Lactobacillus casei (strain ATCC 334)
Length: 404
Description: Predicted ABC-type exoprotein transport system, permease component
Keywords: Complete proteome
(No GO terms)
ID: Q5FIS4
Organism: Lactobacillus acidophilus (strain NCFM)
Length: 403
Description: Putative ABC transporter permease protein
Keywords: Complete proteome
(No GO terms)
1e-40
2e-77
5e-51
4e-53
5e-55
1e-41
1e-178
7e-57
7e-64
1e-35
5e-45
2e-41
5e-67
6e-41
3e-94
1e-47
1e-42
8e-53
1e-56
4e-51
1e-177
4e-35
1e-85
2e-51
4e-44
8e-66
2e-83
1e-54
4e-42
1e-40
3e-37
4e-72
2e-37
2e-82
9e-48
2e-40
1e-146
1e-177
9e-36
1e-88
3e-35
1e-42
2e-37
3e-39
1e-87

Figure 5.7: Candidate group containing the *Lactobacillus helveticus* protein with accession number A8YWH1, created using an E-value threshold of 10^{-34}. If this thesis is being read electronically, the reader may zoom in on the figure so the details can be viewed.

reason for this can be seen by examining Figure 5.7. There appears to be two distinct portions of the graph, with the ten left-most vertices constituting one portion and the six right-most vertices (which include A8YWH1) constituting the other portion. The two portions of the graph are connected by just three edges—two edges connected to the leftmost vertex of the right-hand group, and the other edge connected to the second-to-leftmost vertex of that group. The smallest E-value of any of these three edges is 3×10^{-39}, explaining why, when an E-value threshold of 10^{-39} was used, the A8YWH1 group contained only six proteins—the same six proteins as the right-hand portion of this graph.

Given this, should the group of proteins containing A8YWH1 for E-value thresholds between 10^{-35} and 10^{-38} be considered promising? In other words, is it likely that this group really contains proteins that somehow differentiate the lactobacilli from *P. pentosaceus*? The answer appears to be no. The protein from *P. pentosaceus* exhibited strong homology to the protein from *L. reuteri*, with an E-value of 4×10^{-35}. Furthermore, when a large E-value threshold was chosen, such as 10^{-1}, the graph containing A8YWH1 revealed that the *P. pentosaceus* protein exhibited homology to many of the other *Lactobacillus* proteins from this group, with E-values ranging from 8×10^{-11} to 10^{-33} (graph not shown). It is extremely unlikely that all of these matches would be spurious.

Another example of a group of proteins that was only reported as a candidate group for a very small range of E-value thresholds was represented by the *L. helveticus* protein with accession number A8YTB9. This group, which contained cation efflux proteins, was only reported as a candidate group when an E-value threshold of 10^0 was used. This likely indicates that this is not a true candidate group. At an E-value threshold of 10^0, proteins from all 15 lactobacilli were represented in this group, but no proteins from *P. pentosaceus* were identified as orthologues of any of these proteins. However, when the E-value threshold was decreased to 10^{-1}, there were no proteins from *Lactobacillus sakei* present in the group containing A8YTB9. The graph obtained when an E-value threshold of 10^0 was used is shown in Figure 5.8. Besides illustrating the fact that this is spurious candidate group (the *L. Sakei* proteins in the bottom-right corner are obviously unrelated to the large group of proteins in the middle of the graph), this graph illustrates the value of graphically visualizing orthologous relationships. Described in terms similar to those that were used to describe Figure 5.7, this graph has three portions—the bottom-right portion, described above; the middle portion; and the left-hand portion, which was only in the same connected component as the middle portion by virtue of a bridge through a protein from *Lactobacillus salivarius*, which is strongly homologous (having low E-values) to other proteins from the left-hand portion of the graph, but is weakly related (having relatively high E-values, ranging from 3×10^{-7} to 2.4×10^0) to proteins from the middle portion. Also, note that except for *L. salivarius*, each of the organisms having a protein in the left portion of the graph also has a protein present in the middle portion. These pairs of proteins (a protein from the left portion, and a protein from the same organism in the middle

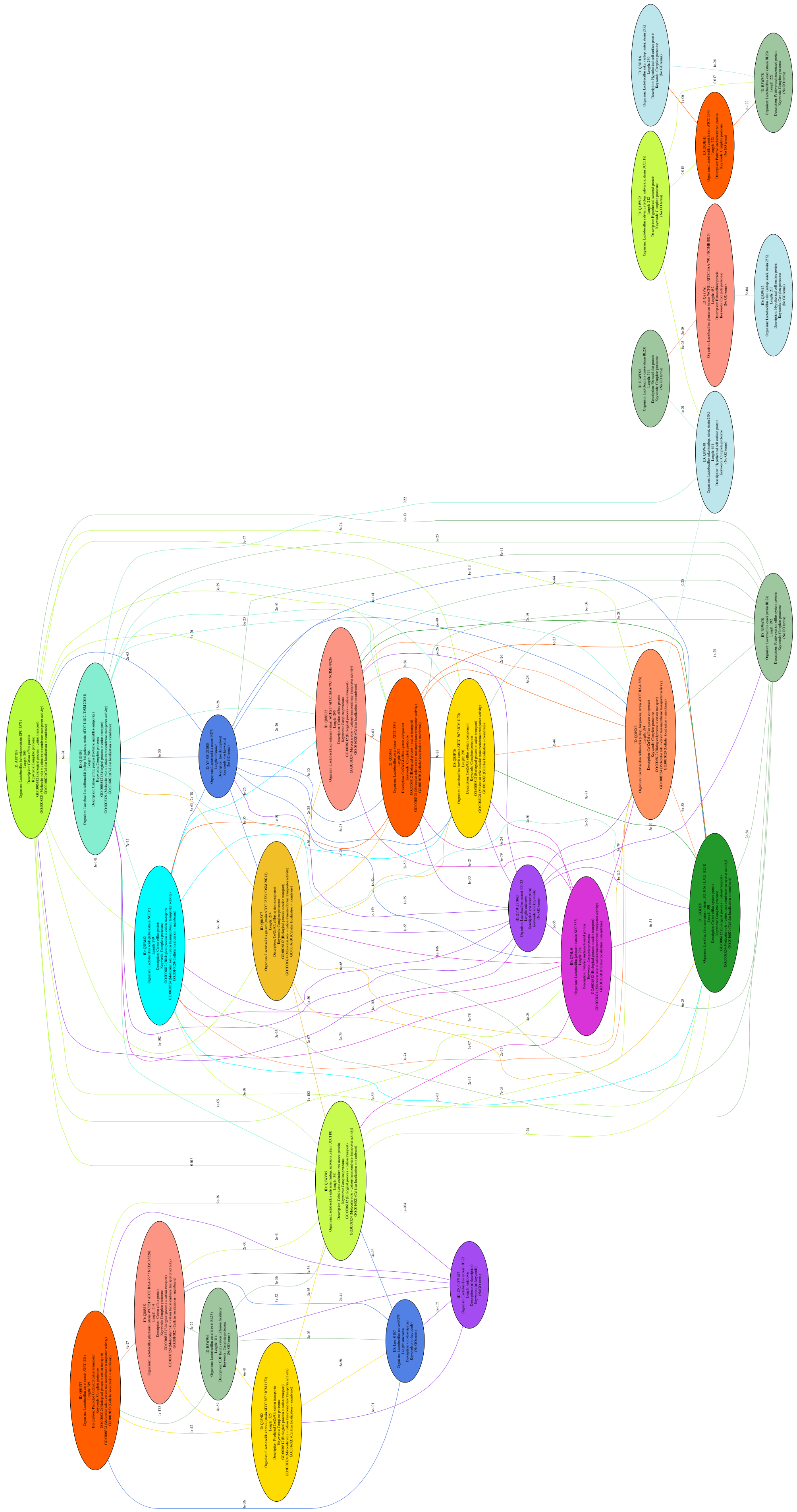

Figure 5.8: Candidate group containing the *Lactobacillus helveticus* protein with accession number A8YTB9, created using an E-value threshold of 10^0. If this thesis is being read electronically, the reader may zoom in on the figure so the details can be viewed.

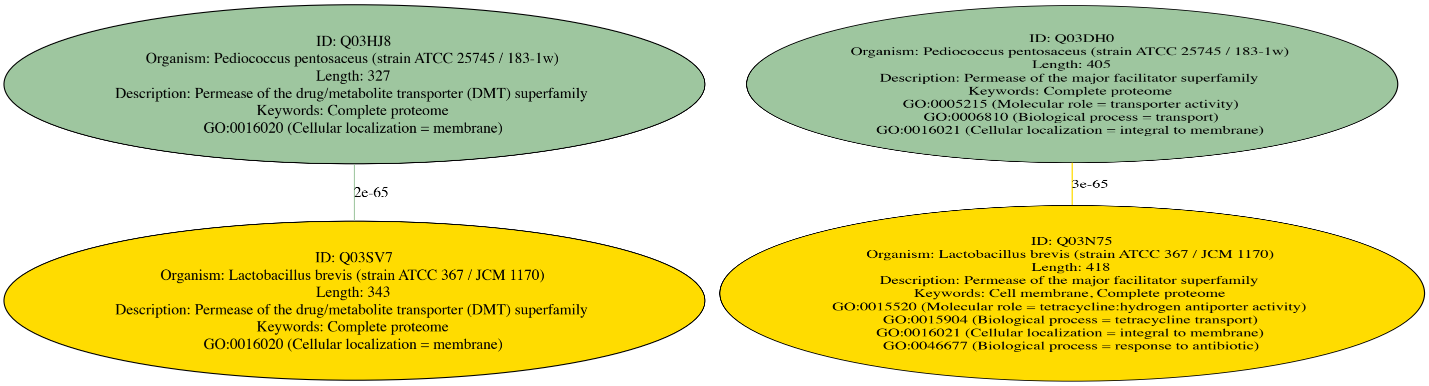
ID: Q03HJ8
Organism: Pediococcus pentosaceus (strain ATCC 25745 / 183-1w)
Length: 327
Description: Permease of the drug/metabolite transporter (DMT) superfamily
Keywords: Complete proteome
GO:0016020 (Cellular localization = membrane)
2e-65
ID: Q03SV7
Organism: Lactobacillus brevis (strain ATCC 367 / JCM 1170)
Length: 343
Description: Permease of the drug/metabolite transporter (DMT) superfamily
Keywords: Complete proteome
GO:0016020 (Cellular localization = membrane)
ID: Q03DH0
Organism: Pediococcus pentosaceus (strain ATCC 25745 / 183-1w)
Length: 405
Description: Permease of the major facilitator superfamily
Keywords: Complete proteome
GO:0005215 (Molecular role = transporter activity)
GO:0006810 (Biological process = transport)
GO:0016021 (Cellular localization = integral to membrane)
3e-65
ID: Q03N75
Organism: Lactobacillus brevis (strain ATCC 367 / JCM 1170)
Length: 418
Description: Permease of the major facilitator superfamily
Keywords: Cell membrane, Complete proteome
GO:0015520 (Molecular role = tetracycline:hydrogen antiporter activity)
GO:0015904 (Biological process = tetracycline transport)
GO:0016021 (Cellular localization = integral to membrane)
GO:0046677 (Biological process = response to antibiotic)

Figure 5.9: Two candidate groups for gatifloxacin resistance.

portion) are likely paralogues—both involved somehow in cation efflux, but perhaps in slightly different capacities. It is also possible that there was once a protein in *L. salivarius* that was orthologous to the proteins in the middle portion, but is no longer present due to gene loss. This example illustrates how inferences can be obtained from these graphs that could not be gleaned using, say, a simple list of the orthologues present in a given group.

5.2.3 Identifying the protein(s) conferring gatifloxacin resistance in LAB

Section 4.2.3 described the methodology used to find proteins that may be responsible for gatifloxacin resistance in the LAB *L. brevis* and *P. pentosaceus*: PSI was used to find proteins that are found in the proteomes of these two organisms, but not in *L. casei* or *L. reuteri* F275, which do not exhibit gatifloxacin resistance. When this analysis was performed, 84 candidate groups were found. Two of the most promising are shown in Figure 5.9. Both groups shown in Figure 5.9 contain proteins annotated as drug transporters, which are not likely to be specific to gatifloxacin, but would nonetheless be good candidates for resistance to any antibiotic. As such, these proteins are ideal candidates for gene knockout experiments, which could determine whether they contribute to gatifloxacin resistance in *L. brevis* and *P. pentosaceus*.

An interesting group of proteins, which was *not* a candidate group, is shown in Figure 5.10. This group contained proteins annotated as quinone reductases; as gatifloxacin belongs to the quinone family of antibiotics, this would initially seem like a promising group; as stated, however, this was not a candidate group, since it contained a protein from *L. casei*, which does not exhibit resistance to gatifloxacin. Thus, this protein cannot (solely) be responsible for gatifloxacin resistance.

There are a number of possible explanations for the existence of this noncandidate group.

- The proteins in this group may actually have nothing to do with gatifloxacin resistance. For instance, they may confer resistance to some other antibiotics in the quinone family, but not gatifloxacin.

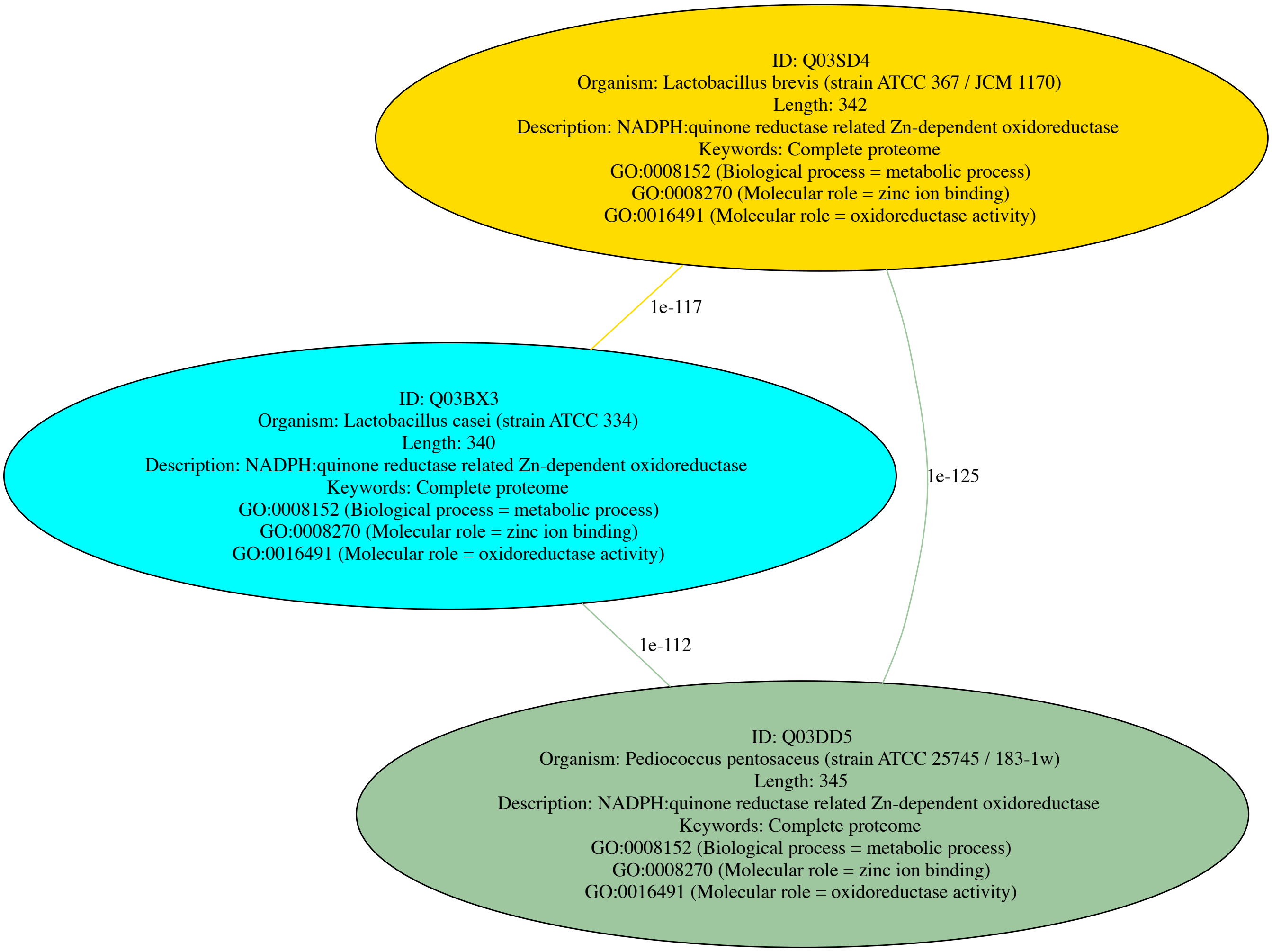
ID: Q03SD4
Organism: Lactobacillus brevis (strain ATCC 367 / JCM 1170)
Length: 342
Description: NADPH:quinone reductase related Zn-dependent oxidoreductase
Keywords: Complete proteome
GO:0008152 (Biological process = metabolic process)
GO:0008270 (Molecular role = zinc ion binding)
GO:0016491 (Molecular role = oxidoreductase activity)
1e-117
1e-125
ID: Q03BX3
Organism: Lactobacillus casei (strain ATCC 334)
Length: 340
Description: NADPH:quinone reductase related Zn-dependent oxidoreductase
Keywords: Complete proteome
GO:0008152 (Biological process = metabolic process)
GO:0008270 (Molecular role = zinc ion binding)
GO:0016491 (Molecular role = oxidoreductase activity)
1e-112
ID: Q03DD5
Organism: Pediococcus pentosaceus (strain ATCC 25745 / 183-1w)
Length: 345
Description: NADPH:quinone reductase related Zn-dependent oxidoreductase
Keywords: Complete proteome
GO:0008152 (Biological process = metabolic process)
GO:0008270 (Molecular role = zinc ion binding)
GO:0016491 (Molecular role = oxidoreductase activity)

Figure 5.10: A noncandidate group for gatifloxacin resistance.

- The proteins from *L. brevis* and *P. pentosaceus* from this group are responsible for gatifloxacin resistance, but the protein is rendered ineffective in *L. casei* due to point mutations in the gene coding for this protein, or due to some other factors unique to *L. casei.*

- Resistance to gatifloxacin is dependent on these proteins interacting with other proteins, which are present in *L. brevis* and *P. pentosaceus*, but not *L. casei.* These other proteins could potentially be found in one or more of the 84 candidate groups.

- Orthologues were detected incorrectly, and the protein from *L. casei* is not actually orthologous to the proteins from *L. brevis* and *P. pentosaceus.* This explanation is unlikely, however, since the E-values were all very small (less than or equal to 10^{-112}).

Based on these data, it remains unclear which protein(s) may be responsible for gatifloxacin resistance in *L. brevis* and *P. pentosaceus*, although the proteins shown in Figure 5.9 are certainly promising candidates. If gene-knockout experiments were performed on the genes that encode the most promising candidate proteins, and none of these proteins were found to confer gatifloxacin resistance, then it would be necessary to resort to other candidate proteins, such as those having completely unknown functions. An example of such a group is given in Figure 5.11. It is also possible that resistance in *L. brevis* and *P. pentosaceus* is primarily due to point mutations in DNA gyrase and DNA topoisomerase IV (the targets of gatifloxacin), rather than the presence of resistance proteins. A final possibility is that *L. brevis* and *P. pentosaceus* do not have the same mechanism of resistance. (See Section 6.4.3 for a more general discussion of the ability or inability of PSI to provide correct results in situations where the cause of the phenotype is not the same in all organisms being examined.)

5.2.4 Determining the relationship between the number of input organisms and the number of candidate proteins

As described in Section 4.2.4, two tests were performed to evaluate the impact of the number of organisms used in a particular PSI comparison on the number of candidate groups returned. The results of these tests are given in the following two sections.

Test 1: Cell shape phenotype

The relationship between the number of lactobacilli involved in a cell shape phenotype comparison and the number of candidate proteins returned is shown in Figure 5.12. This plot shows that the number of candidate proteins decreased rapidly as the number of lactobacilli was increased. When just one *Lactobacillus* isolate was used, the number of candidate proteins was very large (975), making this procedure essentially useless in narrowing down the list of proteins that may be responsible for the difference in cell shape between the lactobacilli and *P. pentosaceus.* However, the

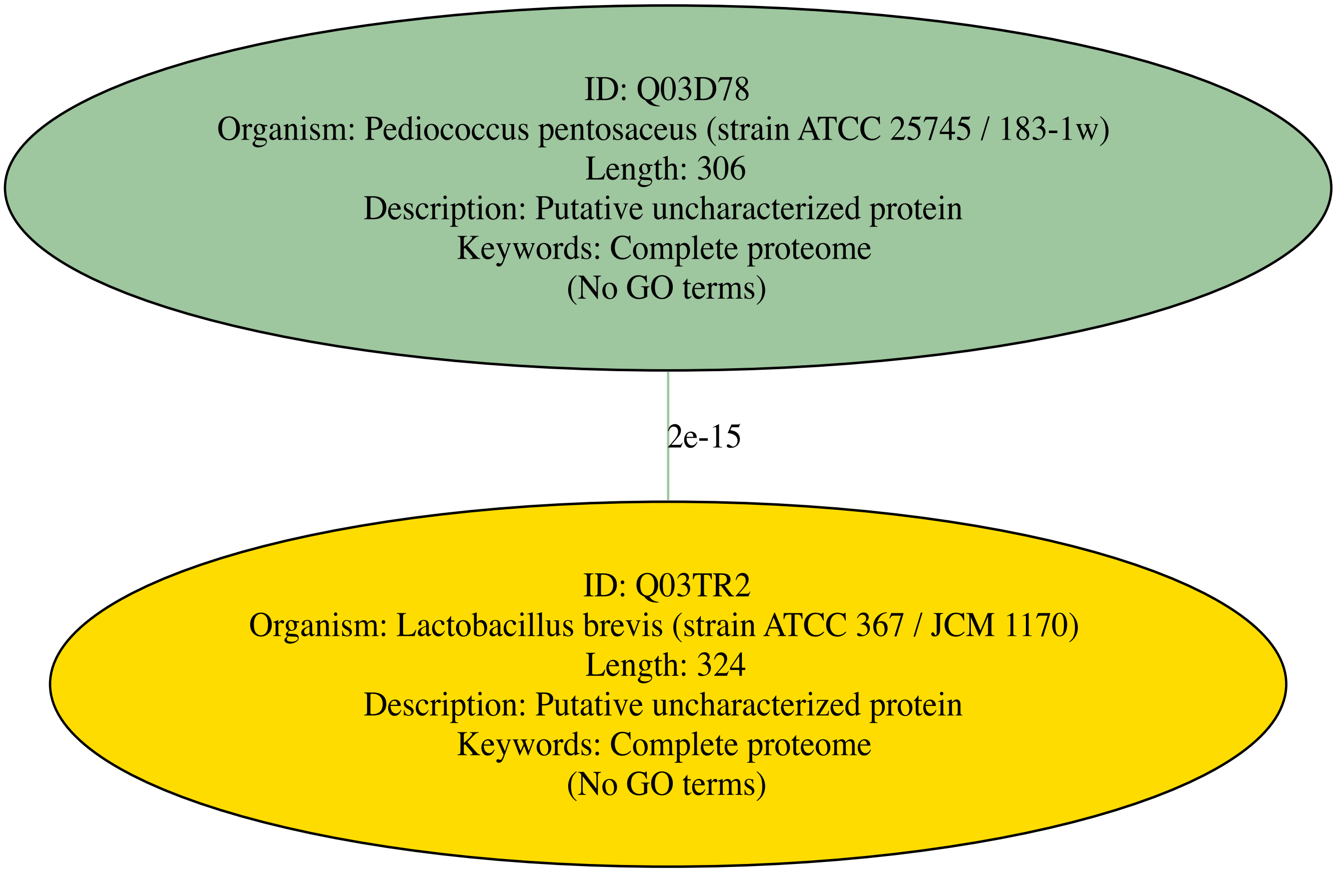
ID: Q03D78
Organism: Pediococcus pentosaceus (strain ATCC 25745 / 183-1w)
Length: 306
Description: Putative uncharacterized protein
Keywords: Complete proteome
(No GO terms)
2e-15
ID: Q03TR2
Organism: Lactobacillus brevis (strain ATCC 367 / JCM 1170)
Length: 324
Description: Putative uncharacterized protein
Keywords: Complete proteome
(No GO terms)

Figure 5.11: A candidate group for gatifloxacin resistance containing uncharacterized proteins.

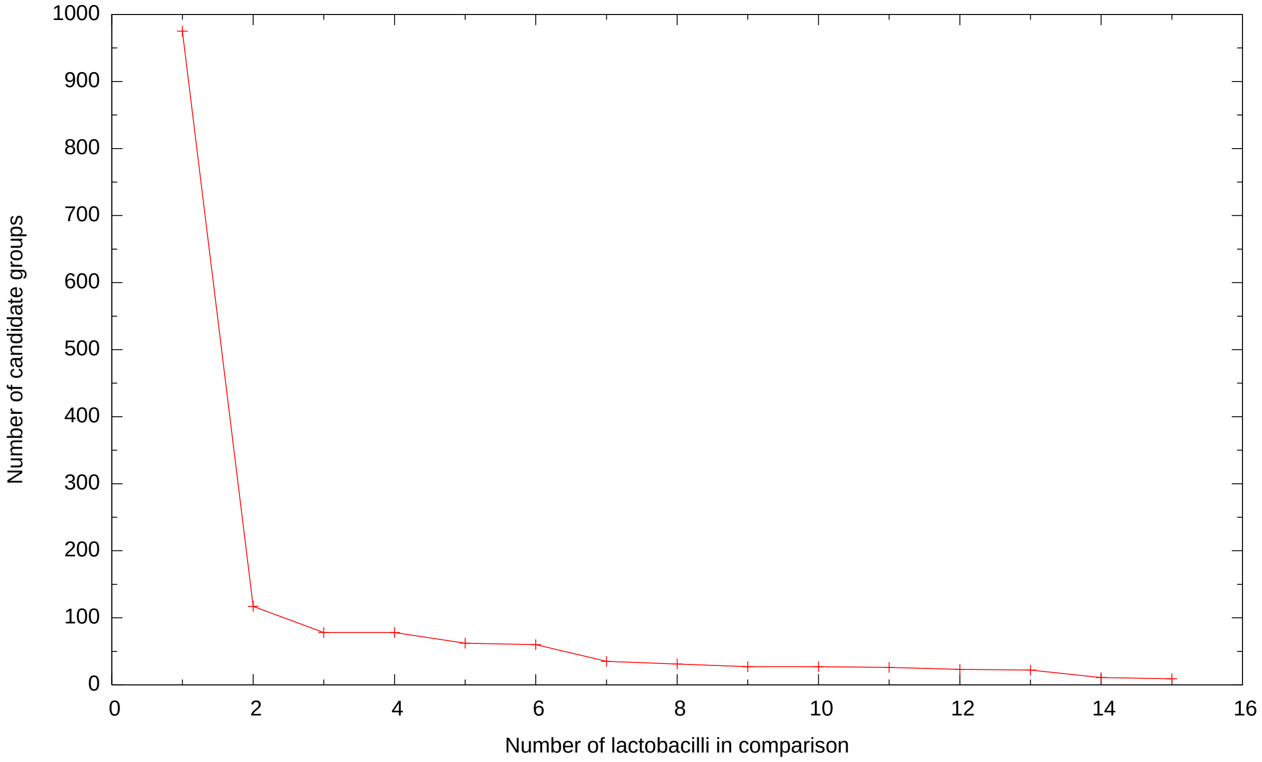

Figure 5.12: Relationship between the number of lactobacilli used in the cell shape phenotype comparison and the number of candidate groups.

number of candidate proteins dropped to just 117 when a second *Lactobacillus* isolate was added, and was less than 40 when seven or more lactobacilli were used. These data suggest that, even with a relatively small number of organisms (less than ten), PSI can successfully generate a list of candidate proteins small enough that they can be examined easily. On the other hand, if laboratory testing was required, then there would still be value in using the greatest number of organisms possible; for instance, using 14 lactobacilli instead of 13 resulted in 11 fewer candidate proteins, which could reduce the laboratory work needed to determine the correct protein(s) significantly.

Test 2: Proteins in *Streptococcus* isolates that are not in *Mycobacterium* isolates

In the second test, it was determined how the number of organisms used affects the number of candidate proteins when PSI was used to find proteins found in all *Streptococcus* isolates, but no *Mycobacterium* isolates. This relationship is depicted in Figure 5.13.

There are two important observations to be made regarding Figure 5.13. First, when the number of mycobacteria in the comparison was held constant, changing the number of streptococci had a very large effect on the number of candidate proteins when the number of streptococci was already small (fewer than five). On the other hand, when the number of streptococci was greater than five, the reduction in the number of candidate proteins when more *Streptococcus* isolates were added

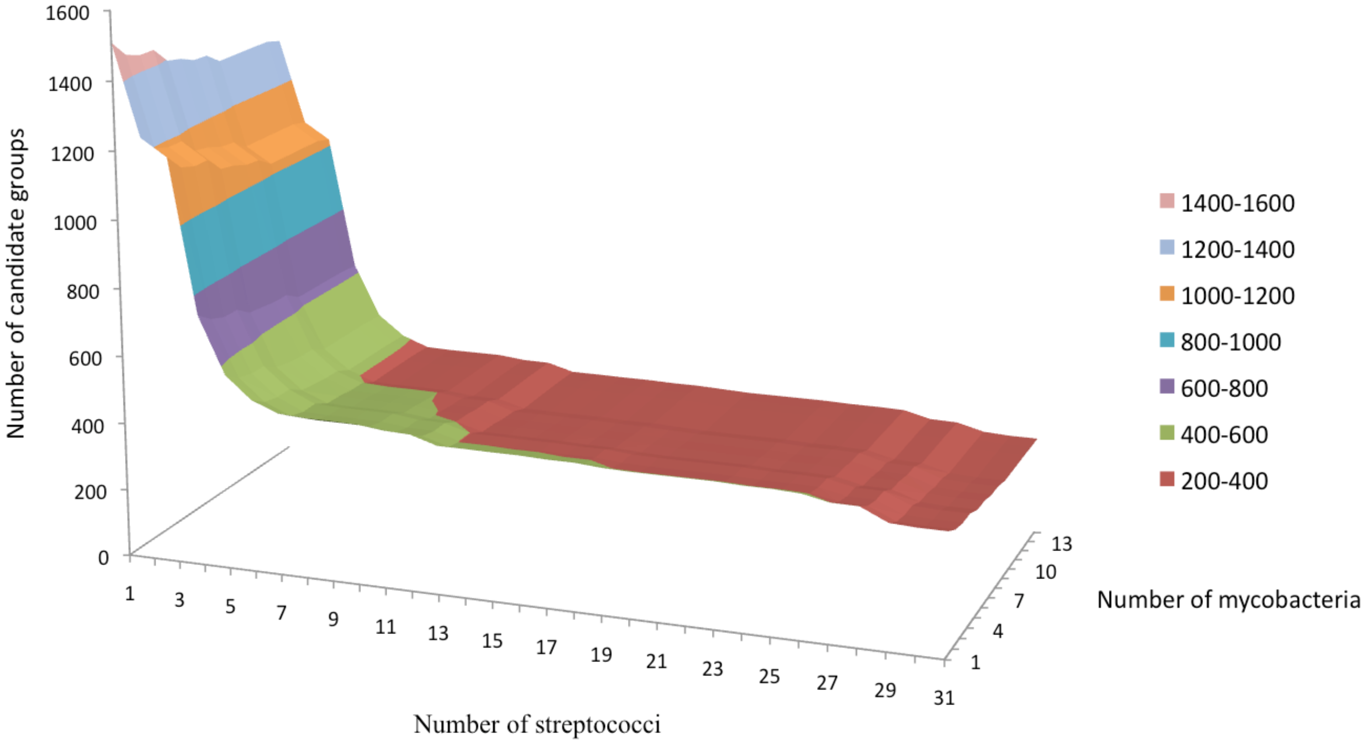

Figure 5.13: Relationship between the number of *Streptococcus* and *Mycobacterium* species used, and the number of candidate proteins, when using PSI to find groups of proteins present in all of the streptococci, but none of the mycobacteria.

was much smaller. Second, when the number of *Streptococcus* isolates was held constant, adding additional *Mycobacterium* isolates had a fairly small impact on the number of candidate proteins. This was irrespective of the number of *Mycobacterium* isolates that had already been added, as well as of the number of *Streptococcus* isolates. However, the impact of adding additional *Mycobacterium* isolates appeared to be somewhat greater when there were few *Streptococcus* isolates compared to when there were many. For instance, when only one *Streptococcus* isolate was used, the number of candidate proteins ranged from 1509 when just one *Mycobacterium* isolate was used to 1319 when all 14 were used. In comparison, when five *Streptococcus* isolates were used, the number of candidate proteins when different numbers of mycobacteria were used ranged from 576 to 468—a smaller difference.

As a consequence of these two observations, Figure 5.13 provides insight into the relative importance of adding additional streptococci compared to adding additional mycobacteria—specifically, that adding additional *Streptococcus* isolates had a greater effect on the number of candidate groups than adding additional *Mycobacterium* isolates, especially when the current number of streptococci was relatively small. To generalize these results, suppose that the user of PSI is trying to find the protein responsible for a particular phenotype, and that the user must perform a literature search to identify organisms that do or do not exhibit the phenotype. These results suggest that the user should concentrate primarily on finding organisms that do exhibit the phenotype, rather than those that do not, since the former seems to reduce the number of candidate groups more quickly than the latter. However, these results show that it is still beneficial to include a few organisms that do not exhibit the phenotype, as well. It should be stressed that generalizing these results may not be justified based on a single experiment, and additional data are necessary to determine whether adding organisms that exhibit the phenotype always has a greater effect on reducing the number of candidate groups than adding organisms that do not exhibit the phenotype. Note that the evolutionary relatedness of the organisms in each category could also play a significant role. If all of the organisms in one of the two sets (those that exhibit the phenotype, and those that do not) were very closely related, then adding more closely related organisms to that same set would likely have little effect on reducing the number of candidate groups; conversely, adding a more distantly related organism would likely have a much greater effect. However, adding an organism that is closely related to those in one set to the *other* set would likely substantially reduce the number of candidate groups, as most proteins in the newly added organism would be orthologous to proteins in the other set, and thus would not be candidate proteins.

5.3 Using PSI for phylogenetics and comparative genomics

In this section, the results of applying PSI to selected phylogenetics and comparative genomics problems are reported. Section 5.3.1 describes the results of determining the effect of the E-value threshold on the number of proteins found to be in one organism, but not a second organism. Section 5.3.2 discusses the use of PSI for creating a large phylogenetic tree based on similarities in the protein content of pairs of bacteria. Finally, Section 5.3.3 gives the results of using PSI to determine how cohesive isolates of the same species are compared to randomly selected sets of isolates from the same genus.

5.3.1 Evaluating the effect of the E-value threshold on numbers of unique proteins

As described in Section 4.3.1, the effect of the E-value threshold on the number of proteins found to be in one organism, but not a second organism, was determined for three different degrees of relatedness: two organisms from the same species; two organisms from the same genus, but different species; and two organisms from different genera. The results of these comparisons are described below.

Intra-species comparisons

A scatterplot illustrating the relationship between the E-value threshold and the number of unique proteins for the intra-species comparisons is given in Figure 5.14. Note that the purpose of this figure is to give three separate examples of how the number of proteins that are reported to be in

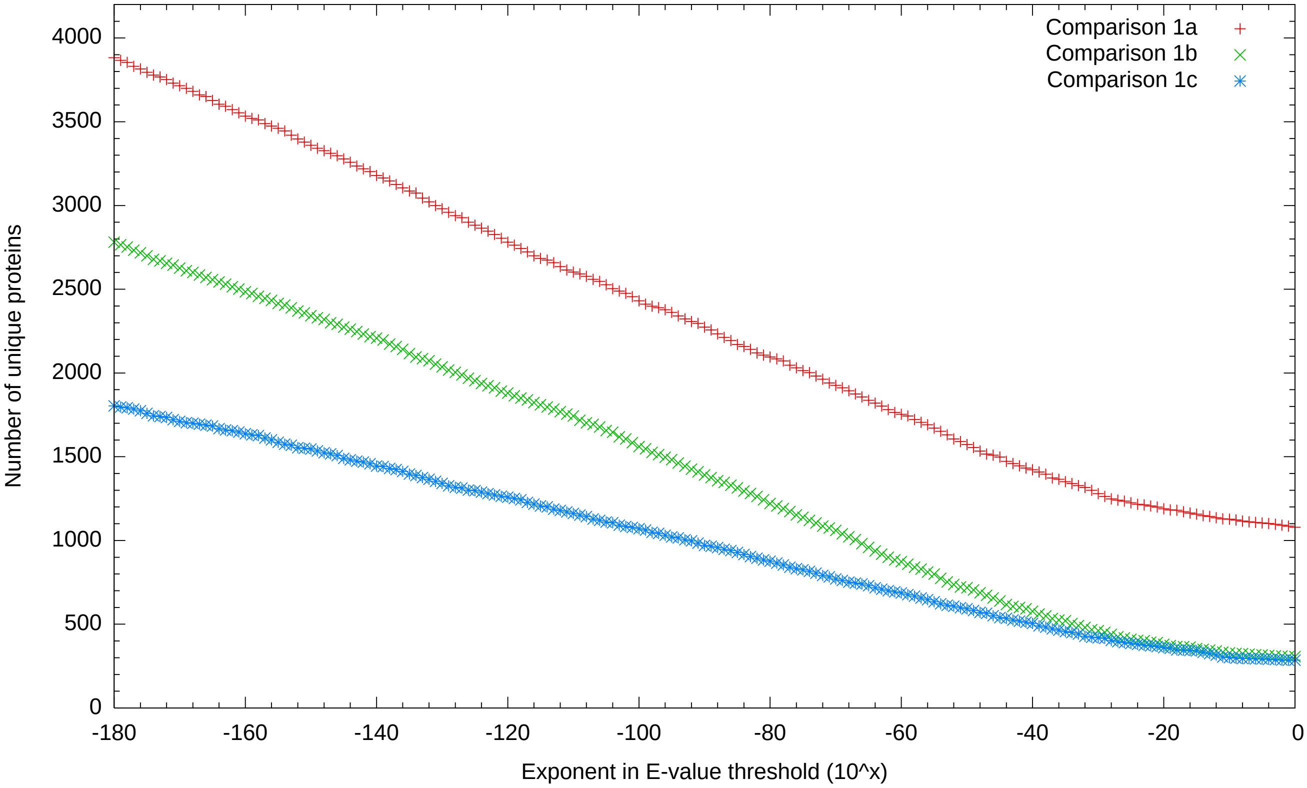

Figure 5.14: The relationship between the E-value threshold and the number of unique proteins reported for pairs of isolates from the same species. See Table 4.1 for the organisms involved in each comparison.

one isolate but not another isolate of the same species varied with the E-value threshold. Thus, the three plots are not meant to be compared with one another. The differences between plots can be partially attributed to the proteome size of each species. For instance, the two *Pseudomonas putida* isolates have very large proteomes (5396 in strain GB-1 and 5313 in strain KT2440), whereas the *Xanthomonas campestris* proteomes are smaller (4239 proteins in strain 8004 and 4410 proteins in strain B100), and the *Staphylococcus aureus* proteomes are smaller yet (2679 proteins in strain COL and 2761 proteins in strain JH1). Thus, one might expect the *Pseudomonas* isolates to have the largest variation, which is consistent with the figure.

As can be seen, the number of unique proteins differed substantially depending on the E-value threshold for all three comparisons. With respect to proteins found in *P. putida* GB-1 but not in *P. putida* KT2440 (comparison 1a), the number of unique proteins reported ranged from 3882 when using an E-value threshold of 10^{-180} to 1075 when using an E-value threshold of 10^{0}. That this range is very wide highlights the importance of choosing an appropriate E-value threshold. A closer look at the plot for *P. putida* revealed that it can be divided into two distinct sections. The first section of the plot ranged from an E-value threshold of 10^{-180} to an E-value threshold of approximately 10^{-30}, in which there was a nearly perfectly linear decrease in the number of unique proteins as the E-value was increased. The second section ranged from E-value thresholds between 10^{-29} and 10^{0}. Like the first section, the number of unique proteins decreased as the E-value was

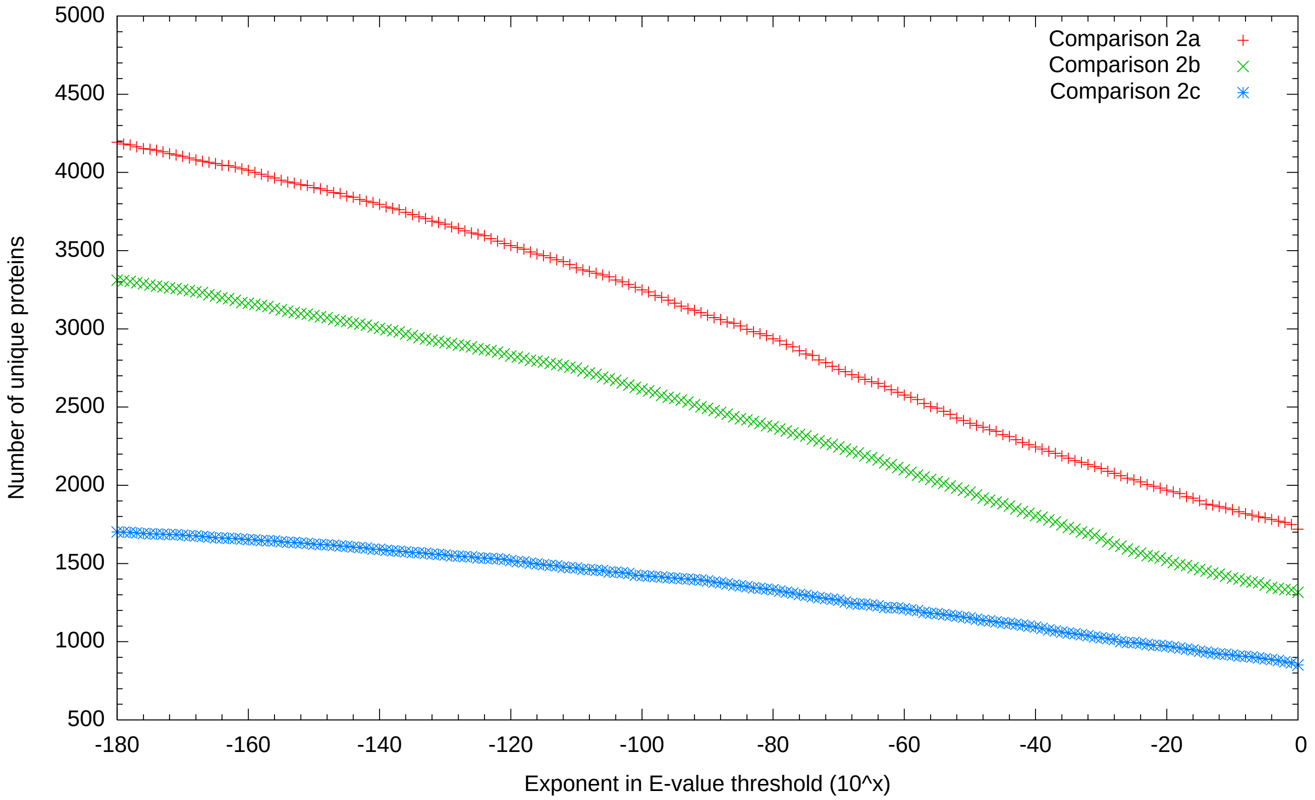

Figure 5.15: The relationship between the E-value threshold and the number of unique proteins reported for pairs of isolates from the same genus, but different species. See Table 4.1 for the organisms involved in each comparison.

increased, although the slope was much smaller. In other words, compared to the first section, increasing the E-value in this region seemed to result in smaller decreases in the number of unique proteins.

The plots for *X. campestris* (comparison 1b) and *S. aureus* (comparison 1c) showed the same general trend as that for *P. putida*, with a nearly constant slope between E-value thresholds of 10^{-180} and 10^{-30}, and a smaller slope between thresholds of 10^{-29} and 10^{0}. Curiously, while there were many more unique proteins in the *X. campestris* comparison than in the *S. aureus* comparison at very stringent E-value thresholds, the number of unique proteins at non-stringent E-value thresholds was nearly the same for these two comparisons. This would seem to reflect the fact that the two *S. aureus* strains are more closely related to each other than the two *X. campestris* strains are to each other.

Inter-species comparisons

A scatterplot depicting the relationship between the E-value threshold and the number of unique proteins for the inter-species (but intra-genus) comparisons is given in Figure 5.15. The plots were similar to those in Figure 5.14; the only exception was that the slope of the line at high E-value thresholds (10^{-29} and above) did not appear to level off as much as it did in the intra-species comparisons.

Inter-genus comparisons

A scatterplot depicting the relationship between the E-value threshold and the number of unique proteins for the three inter-genus comparisons is given in Figure 5.16. All three plots appeared distinctly different than both the intra-species plots and the inter-species plots. The first and second sections of the plot exhibited essentially the opposite trend compared to both the intra-species and inter-species comparisons. The intra-species and inter-species comparisons showed a relatively steep slope between E-value thresholds of approximately 10^{-180} to 10^{-30}, and then a more gradual slope between thresholds of approximately 10^{-29} to 10^{0}. In contrast, the inter-genus plots had a very gradual slope between thresholds of 10^{-180} to 10^{-50}, and then a steeper slope between thresholds of 10^{-49} and 10^{0}.

The relative evolutionary relatedness of the organisms in each comparison type likely accounts for the differences between the intra-species/inter-species comparisons and the inter-genus comparisons. For the intra-species and inter-species comparisons, orthologous proteins would have undergone few mutations compared to the inter-genus comparisons, and such proteins should therefore attain small E-values when one is used as a query to BLAST against a database containing the other protein. Thus, one would expect relatively few changes in the number of unique proteins at larger E-value thresholds, since few true orthologues would attain such large E-values. This trend is reflected in the intra-species and inter-species comparisons shown in Figures 5.14 and 5.15. Conversely, for the inter-genus comparisons, the organisms in each pair were more evolutionarily distant from each other than the pairs in either of the other two comparison types. Given this, orthologues would have undergone many mutations (relative to orthologues in the other two comparison types), resulting in larger E-values. Thus, most orthologues should get relatively large E-values, and therefore the number of unique proteins should change more quickly at large E-value thresholds rather than at small ones. This trend holds for all comparisons shown in Figure 5.16.

The purpose of this section was to assist in choosing an appropriate E-value threshold for analyzing the number of unique proteins in pairs of bacteria. In Section 4.1.6, an equation was presented that related the E-value threshold that should be chosen to the number of proteins in the organisms' proteomes, the number of organisms involved in a particular comparison, and the desired value for the expected number of spurious matches. In that section, it was argued that an E-value threshold of 10^{-13} was appropriate for most comparisons. This value is in good agreement with a value that might be chosen in the basis of Figures 5.14, 5.15, and 5.16. With respect to Figures 5.14, and 5.15, it would seem reasonable to choose an E-value threshold between 10^{-29} and 10^{0}, as the number of unique proteins changed quite slowly as the E-value threshold was varied within this range. It is encouraging that 10^{-13} falls nearly in the middle (logarithmically) of this range. With respect to Figure 5.16, a small E-value (say, less than 10^{-50}) would likely result in most actual orthologues being missed, as most changes in the number of unique proteins

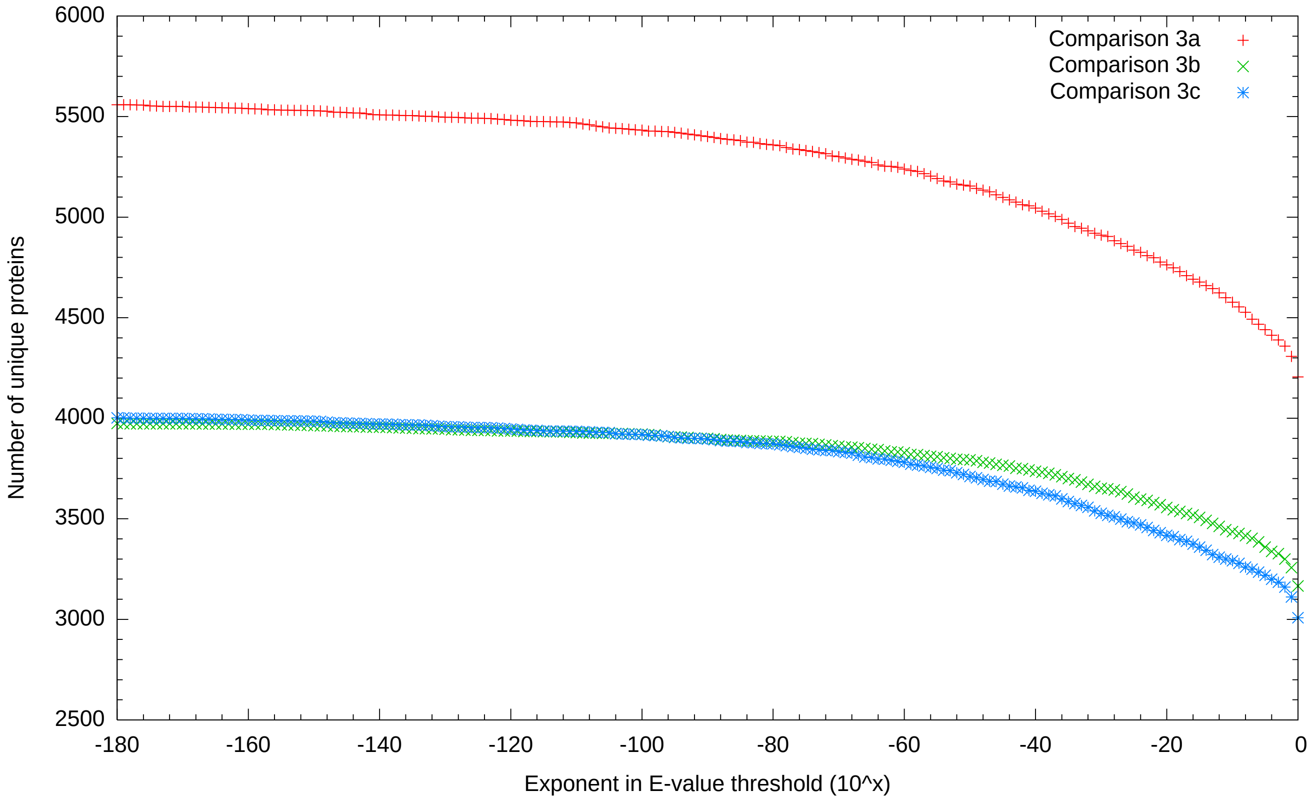

Figure 5.16: The relationship between the E-value threshold and the number of unique proteins reported for pairs of isolates from different genera. See Table 4.1 for the organisms involved in each comparison.

occurred between E-value thresholds of 10^{-49} and 10^{0}. Having narrowed down the range of E-value thresholds that might be chosen based on Figure 5.16, it remains difficult to justify a more precise E-value threshold within this range. Given that a threshold of 10^{-13} is justifiable based on the intra-species and inter-species comparisons, and seems to be within an appropriate range for the inter-genus comparisons, it could be argued that 10^{-13} is an appropriate threshold for any pairwise comparison, especially because it is also supported by the analytical approach from Section 4.1.6. However, it would make interesting future work to investigate the possibility of varying the E-value threshold depending on the evolutionary relatedness of the organisms being considered.

5.3.2 Phylogenetics based on protein content

The phylogenetic tree that was created as described in Section 4.3.3 is shown in Figure 5.17. This tree was created by using as a distance metric the average of the number of proteins found in organism A but not organism B, and vice versa. The tree appeared to be relatively consistent with current taxonomic classifications, with isolates from most genera clustered together. However, there were some exceptions: *Mycobacterium leprae* was isolated from the rest of the mycobacteria, and, in fact, was deemed to be more closely related to the genus *Rickettsia*. Thirteen of the 19 *Clostridium* isolates clustered together, including the *botulinum*, *perfringens*, *tetani*, and *novyi* species. Three other *Clostridium* isolates—*Clostridium phytofermentans*, *Clostridium*

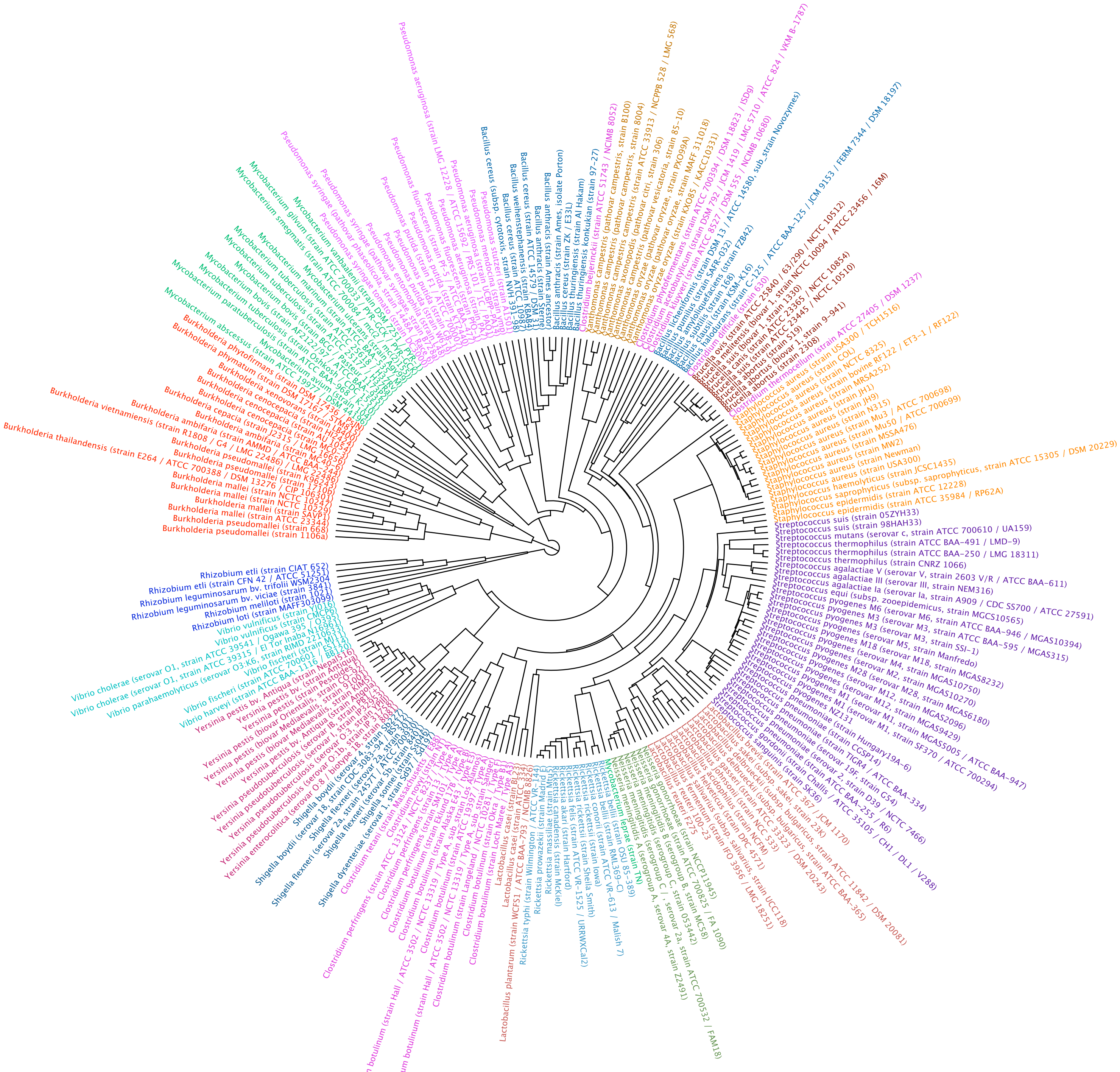
Burkholderia pseudomallei (strain 1106a)
Burkholderia pseudomallei (strain 668)
Burkholderia mallei (strain ATCC 23344)
Burkholderia mallei (strain SAVP1)
Burkholderia mallei (strain NCTC 10229)
Burkholderia mallei (strain NCTC 10247)
Burkholderia thailandensis (strain E264 / ATCC 700388 / DSM 13276 / CIP 106301)
Burkholderia pseudomallei (strain 1710b)
Burkholderia pseudomallei (strain K96243)
Burkholderia vietnamiensis (strain R1808 / G4 / LMG 22486)
Burkholderia ambifaria (strain AMMD / ATCC BAA-244)
Burkholderia ambifaria (strain MC40-6)
Burkholderia cepacia (strain J2315 / LMG 16656)
Burkholderia cenocepacia (strain MC0-3)
Burkholderia cenocepacia (strain AU 1054)
Burkholderia cenocepacia (strain HI2424)
Burkholderia xenovorans (strain LB400)
Burkholderia phymatum (strain DSM 17167 / STM815)
Burkholderia phytofirmans (strain DSM 17436 / PsJN)
Mycobacterium abscessus (strain ATCC 19977 / DSM 44196)
Mycobacterium avium (strain 104)
Mycobacterium paratuberculosis (strain ATCC BAA-968 / K-10)
Mycobacterium tuberculosis (strain CDC 1551 / Oshkosh)
Mycobacterium bovis (strain AF2122/97)
Mycobacterium bovis (strain BCG / Pasteur 1173P2)
Mycobacterium tuberculosis (strain ATCC 25177 / H37Ra)
Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv)
Mycobacterium marinum (strain ATCC BAA-535 / M)
Mycobacterium ulcerans (strain Agy99)
Mycobacterium smegmatis (strain ATCC 700084 / mc(2)155)
Mycobacterium gilvum (strain ATCC 700033 / PYR-GCK)
Mycobacterium vanbaalenii (strain DSM 7251 / PYR-1)
Pseudomonas syringae (pathovar phaseolicola, strain 1448A / Race 6)
Pseudomonas syringae tomato (strain DC3000)
Pseudomonas syringae (pathovar syringae, strain B728a)
Pseudomonas entomophila (strain L48)
Pseudomonas putida (strain W619)
Pseudomonas putida (strain GB-1)
Pseudomonas putida (strain F1 / ATCC 700007)
Pseudomonas fluorescens (strain Pf-5 / ATCC BAA-477)
Pseudomonas putida (strain KT2440)
Pseudomonas fluorescens (strain PfO-1)
Pseudomonas aeruginosa (strain LMG 12228 / ATCC 15692 / PRS 101 / 1C / PAO1)
Pseudomonas aeruginosa (strain PA7)
Pseudomonas aeruginosa (strain UCBPP-PA14)
Pseudomonas mendocina (strain ymp)
Pseudomonas stutzeri (strain A1501)
Bacillus cereus (subsp. cytotoxis, strain NVH 391-98)
Bacillus cereus (strain ATCC 10987)
Bacillus weihenstephanensis (strain KBAB4)
Bacillus cereus (strain ATCC 14579 / DSM 31)
Bacillus anthracis (strain Sterne)
Bacillus anthracis (strain Ames ancestor)
Bacillus anthracis (strain Ames, isolate Porton)
Bacillus cereus (strain ZK / E33L)
Bacillus thuringiensis (strain Al Hakam)
Bacillus thuringiensis konkukian (strain 97-27)
Clostridium beijerinckii (strain ATCC 51743 / NCIMB 8052)
Xanthomonas campestris (pathovar campestris, strain B100)
Xanthomonas campestris (pathovar campestris, strain 8004)
Xanthomonas campestris campestris (strain ATCC 33913 / NCPPB 528 / LMG 568)
Xanthomonas axonopodis (pathovar citri, strain 306)
Xanthomonas campestris (pathovar vesicatoria, strain 85-10)
Xanthomonas oryzae (pathovar oryzae, strain PXO99A)
Xanthomonas oryzae (pathovar oryzae, strain MAFF 311018)
Xanthomonas oryzae oryzae (strain KXO85 / KACC10331)
Clostridium phytofermentans (strain ATCC 700394 / DSM 18823 / ISDg)
Clostridium acetobutylicum (strain DSM 792 / JCM 1419 / LMG 5710 / ATCC 824 / VKM B-1787)
Clostridium kluyveri (strain ATCC 8527 / DSM 555 / NCIMB 10680)
Bacillus licheniformis (strain DSM 13 / ATCC 14580, sub_strain Novozymes)
Bacillus pumilus (strain SAFR-032)
Bacillus amyloliquefaciens (strain FZB42)
Bacillus subtilis (strain 168)
Bacillus clausii (strain KSM-K16)
Bacillus halodurans (strain C-125 / ATCC BAA-125 / JCM 9153 / FERM 7344 / DSM 18197)
Clostridium difficile (strain 630)
Brucella ovis (strain ATCC 25840 / 63/290 / NCTC 10512)
Brucella melitensis (biovar 1, strain NCTC 10094 / ATCC 23456 / 16M)
Brucella suis (biovar 1, strain 1330)
Brucella canis (strain ATCC 23365 / NCTC 10854)
Brucella suis (strain ATCC 23445 / NCTC 10510)
Brucella abortus (strain S19)
Brucella abortus (biovar 1, strain 9-941)
Brucella abortus (strain 2308)
Clostridium thermocellum (strain ATCC 27405 / DSM 1237)
Staphylococcus aureus (strain USA300 / TCH1516)
Staphylococcus aureus (strain COL)
Staphylococcus aureus (strain bovine RF122 / ET3-1 / RF122)
Staphylococcus aureus (strain NCTC 8325)
Staphylococcus aureus (strain MRSA252)
Staphylococcus aureus (strain JH1)
Staphylococcus aureus (strain JH9)
Staphylococcus aureus (strain N315)
Staphylococcus aureus (strain Mu3 / ATCC 700698)
Staphylococcus aureus (strain Mu50 / ATCC 700699)
Staphylococcus aureus (strain MSSA476)
Staphylococcus aureus (strain MW2)
Staphylococcus aureus (strain Newman)
Staphylococcus aureus (strain USA300)
Staphylococcus haemolyticus (strain JCSC1435)
Staphylococcus saprophyticus (subsp. saprophyticus, strain ATCC 15305 / DSM 20229)
Staphylococcus epidermidis (strain ATCC 12228)
Staphylococcus epidermidis (strain ATCC 35984 / RP62A)
Streptococcus suis (strain 05ZYH33)
Streptococcus suis (strain 98HAH33)
Streptococcus mutans (serovar c, strain ATCC 700610 / UA159)
Streptococcus thermophilus (strain ATCC BAA-491 / LMD-9)
Streptococcus thermophilus (strain ATCC BAA-250 / LMG 18311)
Streptococcus thermophilus (strain CNRZ 1066)
Streptococcus agalactiae V (serovar V, strain 2603 V/R / ATCC BAA-611)
Streptococcus agalactiae III (serovar III, strain NEM316)
Streptococcus agalactiae Ia (serovar Ia, strain A909 / CDC SS700 / ATCC 27591)
Streptococcus equi (subsp. zooepidemicus, strain MGCS10565)
Streptococcus pyogenes M6 (serovar M6, strain ATCC BAA-946 / MGAS10394)
Streptococcus pyogenes M3 (serovar M3, strain ATCC BAA-595 / MGAS315)
Streptococcus pyogenes M3 (serovar M3, strain SSI-1)
Streptococcus pyogenes (serovar M5, strain Manfredo)
Streptococcus pyogenes M18 (serovar M18, strain MGAS8232)
Streptococcus pyogenes (serovar M4, strain MGAS10750)
Streptococcus pyogenes (serovar M2, strain MGAS10270)
Streptococcus pyogenes M28 (serovar M28, strain MGAS6180)
Streptococcus pyogenes (serovar M12, strain MGAS2096)
Streptococcus pyogenes (serovar M12, strain MGAS9429)
Streptococcus pyogenes M1 (serovar M1, strain MGAS5005 / ATCC BAA-947)
Streptococcus pyogenes M1 (serovar M1, strain SF370 / ATCC 700294)
Streptococcus pyogenes NZ131
Streptococcus pneumoniae (strain Hungary19A-6)
Streptococcus pneumoniae (strain CGSP14)
Streptococcus pneumoniae (strain TIGR4 / ATCC BAA-334)
Streptococcus pneumoniae (serovar 19F, strain G54)
Streptococcus pneumoniae (serovar 2, strain D39 / NCTC 7466)
Streptococcus pneumoniae (strain ATCC BAA-255 / R6)
Streptococcus gordonii (strain Challis / ATCC 35105 / CH1 / DL1 / V288)
Streptococcus sanguinis (strain SK36)
Lactobacillus brevis (strain ATCC 367 / JCM 1170)
Lactobacillus sakei (subsp. sakei, strain 23K)
Lactobacillus delbrueckii (subsp. bulgaricus, strain ATCC 11842 / DSM 20081)
Lactobacillus delbrueckii (subsp. bulgaricus, strain ATCC BAA-365)
Lactobacillus gasseri (strain ATCC 33323 / DSM 20243)
Lactobacillus johnsonii (strain NCC 533)
Lactobacillus acidophilus (strain NCFM)
Lactobacillus helveticus (strain DPC 4571)
Lactobacillus salivarius (subsp. salivarius, strain UCC118)
Lactobacillus fermentum (strain IFO 3956 / LMG 18251)
Lactobacillus reuteri 100-23
Lactobacillus reuteri F275
Neisseria gonorrhoeae (strain NCCP11945)
Neisseria gonorrhoeae (strain ATCC 700825 / FA 1090)
Neisseria meningitidis B (serogroup B, strain MC58)
Neisseria meningitidis (serogroup C, strain 053442)
Neisseria meningitidis (serogroup C / , serovar 2a, strain ATCC 700532 / FAM18)
Neisseria meningitidis A (serogroup A, serovar 4A, strain Z2491)
Mycobacterium leprae (strain TN)
Rickettsia bellii (strain OSU 85-389)
Rickettsia bellii (strain RML369-C)
Rickettsia conorii (strain ATCC VR-613 / Malish 7)
Rickettsia rickettsii (strain Iowa)
Rickettsia rickettsii (strain Sheila Smith)
Rickettsia felis (strain ATCC VR-1525 / URRWXCal2)
Rickettsia akari (strain Hartford)
Rickettsia canadensis (strain McKiel)
Rickettsia massiliae (strain Mtu5)
Rickettsia prowazekii (strain Madrid E)
Rickettsia typhi (strain Wilmington / ATCC VR-144)
Lactobacillus plantarum (strain WCFS1 / ATCC BAA-793 / NCIMB 8826)
Lactobacillus casei (strain ATCC 334)
Lactobacillus casei (strain BL23)
Clostridium botulinum (strain Loch Maree / Type A3)
Clostridium botulinum (strain Okra / Type B1)
Clostridium botulinum (strain Langeland / NCTC 10281 / Type F)
Clostridium botulinum (strain Hall / ATCC 3502 / NCTC 13319 / Type A, sub_strain Sanger)
Clostridium botulinum (strain ATCC 19397 / Type A)
Clostridium botulinum (strain Hall / ATCC 3502 / NCTC 13319 / Type A, sub_strain Los Alamos)
Clostridium botulinum (strain Alaska E43 / type E3)
Clostridium botulinum (strain Eklund 17B / type B)
Clostridium perfringens (strain SM101 / Type A)
Clostridium perfringens (strain 13 / Type A)
Clostridium novyi (strain NT)
Clostridium perfringens (strain ATCC 13124 / NCTC 8237 / Type A)
Clostridium tetani (strain Massachusetts / E88)
Shigella dysenteriae (serovar 1, strain Sd197)
Shigella sonnei (strain Ss046)
Shigella flexneri (serovar 5b, strain 8401)
Shigella flexneri (serovar 2a, strain 2457T / ATCC 700930)
Shigella flexneri (serovar 2a, strain 301)
Shigella boydii (serovar 18, strain CDC 3083-94 / BS512)
Shigella boydii (serovar 4, strain Sb227)
Yersinia enterocolitica (serovar O:8 / biotype 1B, strain 8081)
Yersinia pseudotuberculosis (serovar O:1b, strain IP 31758)
Yersinia pseudotuberculosis (serovar O:3, strain YPIII)
Yersinia pseudotuberculosis (serovar I, strain IP32953)
Yersinia pseudotuberculosis (serovar IB, strain PB1/+)
Yersinia pestis bv. Antiqua (strain Angola)
Yersinia pestis (biovar Mediaevalis, strain KIM5)
Yersinia pestis (biovar Mediaevalis, strain 91001)
Yersinia pestis (biovar Orientalis, strain CO-92)
Yersinia pestis (strain Pestoides F)
Yersinia pestis bv. (strain Antiqua)
Yersinia pestis bv. Antiqua (strain Nepal516)
Vibrio harveyi (strain ATCC BAA-1116 / BB120)
Vibrio fischeri (strain ATCC 700601 / ES114)
Vibrio fischeri (strain MJ11)
Vibrio parahaemolyticus (serovar O3:K6, strain RIMD 2210633)
Vibrio cholerae (serovar O1, strain ATCC 39315 / El Tor Inaba N16961)
Vibrio cholerae (serovar O1, strain ATCC 39541 / Ogawa 395 / O395)
Vibrio vulnificus (strain CMCP6)
Vibrio vulnificus (strain YJ016)
Rhizobium loti (strain MAFF303099)
Rhizobium meliloti (strain 1021)
Rhizobium leguminosarum bv. viciae (strain 3841)
Rhizobium leguminosarum bv. trifolii WSM2304
Rhizobium etli (strain CFN 42 / ATCC 51251)
Rhizobium etli (strain CIAT 652)

Figure 5.17: Phylogenetic tree created using the AUP distance metric, and UPGMA as the linkage method. Organisms from the same species have identical colours. If this thesis is being read electronically, the reader may zoom in on the figure so the details can be viewed.

acetobutylicum, and *Clostridium kluyveri*—appeared together near isolates of the *Bacillus* genus. The remaining *Clostridium* isolates—*Clostridium beijerinckii*, *Clostridium difficile*, and *Clostridium thermocellum*—did not appear to be closely related to any of the other isolates, suggesting that they may be more closely related to taxonomic groups not included in this tree. The final genus that was split up was *Lactobacillus*, with *L. plantarum* and the two *L. casei* isolates being separated from the other lactobacilli.

5.3.3 Evaluating taxonomic classifications by determining how well species are clustered based on protein content

Results from the analysis described in Section 4.3.4 are given in Tables 5.3 and 5.4. From a protein content perspective, the classification of a set of organisms into a single species could be described as "good" if two criteria are met: the organisms are very similar to each other (i.e., have a large core proteome), and are distinct from other organisms (i.e., have many proteins not found in other organisms of the same genus). These two criteria were investigated by comparing the core proteome and the number of unique proteins in a given species to randomly-generated sets of isolates from the same genus. This was done for each species from the genera listed in Table 4.2 that had two or more isolates sequenced.

As an example of reading Tables 5.3 and 5.4, consider the first row of Table 5.3, which contains *B. anthracis*. There were 4941 proteins found in all three sequenced isolates of *B. anthracis*. However, when sets of three *Bacillus* isolates were randomly chosen as described in Section 4.3.4, the average core proteome size was just 2123. According to a two-tailed t-test, the P-value for this comparison was less than 0.001, indicating that the difference in core proteome size between the three *B. anthracis* isolates, and randomly chosen sets of three *Bacillus* isolates, was highly statistically significant. In fact, none of the 25 randomly-generated sets contained a larger core proteome than the set of *B. anthracis* isolates. Therefore, *B. anthracis* satisfied the first criterion specified in Section 4.3.4—the three *B. anthracis* isolates had more similar protein content than randomly-chosen sets of three *Bacillus* isolates. *B. anthracis* also satisfied the second criterion, which stated that species should be distinct from other isolates of the same genus. Table 5.3 shows that the *B. anthracis* isolates contained 168 proteins not found in any other *Bacillus* isolate, compared to an average of just one unique protein for the 25 randomly-generated sets ($P < 0.001$). None of the 25 randomly-generated sets contained more unique proteins than the three *B. anthracis* isolates. Overall, the fact that *B. anthracis* satisfied both criteria suggests that its current taxonomic classification is sound.

As another example, consider *Rhizobium leguminosarum*. There were 3678 proteins in its core proteome, compared to an average of 3386 for randomly selected sets of two *Rhizobium* isolates. This difference was statistically significant; however, two of the 13 random sets of two *Rhizobium* isolates

Table 5.3: Results of experiments concerning clustering of protein content. The meanings of the column headings, as well as other details regarding the table, can be found below the table. See Table 5.4 for the continuation of this table.

Species		Core proteomes				Unique proteins			
Species name	$\mathbf{N_I}$	$\mathbf{N^C_A}$	$\mathbf{N^C_R}$	$\mathbf{P^C}$	$\mathbf{N^C_>}$	$\mathbf{N^U_A}$	$\mathbf{N^U_R}$	$\mathbf{P^U}$	$\mathbf{N^U_>}$
Bacillus anthracis	3	4941	2123	**	0/25	168	1	**	0/25
Bacillus cereus	4	2881	1840	**	0/25	2	0	–	0/25
Bacillus thuringiensis	2	4255	2864	**	5/25	4	7	n.s.	7/25
Brucella abortus	3	2699	2603	**	6/25	2	1	*	4/25
Brucella suis	2	3025	2760	**	2/24	5	4	n.s.	5/24
Burkholderia ambifaria	2	5609	3798	**	1/25	198	17	**	0/25
Burkholderia cenocepacia	3	5908	3352	**	0/25	168	0	**	0/25
Burkholderia mallei	4	3623	3086	**	1/25	18	0	–	0/25
Burkholderia pseudomallei	4	4972	3086	**	0/25	45	0	–	0/25
Clostridium botulinum	8	1514	763	**	0/25	10	0	–	0/25
Clostridium perfringens	3	2110	1085	**	0/25	298	0	**	0/25
Lactobacillus casei	2	2355	959	**	0/25	593	5	**	0/25
Lactobacillus delbrueckii	2	1372	959	**	0/25	222	5	**	0/25
Lactobacillus reuteri	2	1402	959	**	0/25	120	5	**	0/25
Mycobacterium bovis	2	3822	2577	**	1/25	36	38	n.s.	3/25
Mycobacterium tuberculosis	3	3724	2118	**	0/25	26	17	n.s.	3/25
Neisseria gonorrhoeae	2	1795	1560	**	0/8	229	3	**	0/8
Neisseria meningitidis	4	1547	1426	**	0/14	75	4	**	0/14

Column heading abbreviations are as follows: N_I, number of sequenced isolates from the species in the first column; N^C_A, actual size of the core proteome of all the sequenced isolates of that species; N^C_R, average core proteome size of the randomly-generated sets; P^C, probability that the average core proteome size of the randomly-generated sets is different than the actual size of the core proteome of the sequenced isolates of this species; $N^C_>$, number of random sets (out of the total number of random sets) having a core proteome larger than that of the species from the first column; N^U_A, actual number of proteins found in all isolates of the species from the first column, but no other isolates from the same genus ("unique proteins"); N^U_R, average number of unique proteins for the randomly-generated sets; P^U, probability that the average number of unique proteins in the randomly-generated sets is different than the actual number of unique proteins in the sequenced isolates of this species; $N^U_>$, number of random sets (out of the total number of random sets) having more unique proteins than the species from the first column.

In some cases, all of the random sets corresponding to a particular species had zero unique proteins. No P-value could be computed for these because the standard deviation of these values was zero. In these situations, the P^U column contains a dash character (–). The averages in both column N^C_R and column N^U_R are rounded to the nearest whole number. For certain rows, column N^U_R shows a value of 0; in some cases, this value is exact, while in other situations, it is due to rounding. If due to rounding, then the standard deviation of the random sets is non-zero, and column P^U contains a P-value. For columns P^C and P^U, "n.s." means "not significant", a single asterisk indicates a P-value of less than 0.05, and a double asterisk indicates a P-value of less than 0.001.

Table 5.4: Results of experiments concerning clustering of protein content (continued). For the meanings of each column, see Figure 5.3.

Species		Core proteomes				Unique proteins			
Species name	$\mathbf{N_I}$	$\mathbf{N_A^C}$	$\mathbf{N_R^C}$	$\mathbf{P^C}$	$\mathbf{N_>^C}$	$\mathbf{N_A^U}$	$\mathbf{N_R^U}$	$\mathbf{P^U}$	$\mathbf{N_>^U}$
Pseudomonas aeruginosa	3	4959	2877	**	0/25	571	1	**	0/25
Pseudomonas fluorescens	2	4206	3199	**	0/25	142	6	**	0/25
Pseudomonas putida	4	3799	2592	**	0/25	69	0	**	0/25
Pseudomonas syringae	3	3894	2877	**	0/25	290	1	**	0/25
Rhizobium etli	2	4700	3386	**	0/13	251	88	**	1/13
Rhizobium leguminosarum	2	3678	3386	*	2/13	99	88	n.s.	5/13
Rickettsia bellii	2	1277	850	**	0/25	219	1	**	0/25
Rickettsia rickettsii	2	1221	850	**	0/25	93	1	**	0/25
Shigella boydii	2	3170	2989	**	2/17	95	12	**	0/17
Shigella flexneri	3	3255	2770	**	0/25	130	6	**	0/25
Staphylococcus aureus	14	1917	1486	**	0/25	157	0	**	0/25
Staphylococcus epidermidis	2	2080	1798	**	0/25	131	0	**	0/25
Streptococcus agalactiae	3	1688	1019	**	0/25	156	0	–	0/25
Streptococcus pneumoniae	6	1543	922	**	0/25	150	0	–	0/25
Streptococcus pyogenes	13	1348	811	**	0/25	49	0	–	0/25
Streptococcus suis	2	1971	1087	**	0/25	336	0	**	0/25
Streptococcus thermophilus	3	1359	1019	**	0/25	145	0	–	0/25
Vibrio cholerae	2	3384	2764	**	1/25	425	20	**	0/25
Vibrio fischeri	2	3380	2764	**	1/25	447	20	**	0/25
Vibrio vulnificus	2	3882	2764	**	0/25	321	20	**	0/25
Xanthomonas campestris	4	3376	2818	**	0/25	49	4	**	0/25
Xanthomonas oryzae	3	3276	2915	**	5/25	299	0	**	0/25
Yersinia pestis	7	2986	2717	**	4/25	21	0	**	0/25
Yersinia pseudotuberculosis	4	3424	3003	**	0/25	21	0	**	0/25

did have core proteomes containing more than 3678 proteins. The two *Rhizobium leguminosarum* isolates had 99 proteins found in both of those isolates, but in no other *Rhizobium* isolate. However, the average for the random sets was 88—a difference that was not statistically significant. Moreover, five of the 13 random sets had more than 99 unique proteins. This may indicate that the taxonomy of *Rhizobium leguminosarum*, or of the entire *Rhizobium* genus, may need to be revised.

Tables 5.3 and 5.4 show that the isolates of most species had both larger core proteomes and a greater number of unique proteins than the isolates in the corresponding randomly-generated sets. However, there were some exceptions. For instance, *Bacillus cereus* had a much larger core proteome than the randomly generated sets, but had just two unique proteins—greater than the average number of unique proteins in the randomly-generated sets, none of which had any unique proteins, but much less than the number of unique proteins possessed by other species having four (or more) sequenced isolates. *Burkholderia mallei*, also with four sequenced isolates, had 18 unique proteins; other species having more than four sequenced isolates were: *Clostridium botulinum* (10 unique proteins), *N. meningitidis* (75), *P. putida* (69), *S. aureus* (157), *Streptococcus pneumoniae* (150), *S. pyogenes* (49), *X. campestris* (49), *Yersinia pestis* (21), and *Yersinia pseudotuberculosis* (21). This may indicate that the taxonomic classification of *B. cereus* should be reexamined. Another example was *Bacillus thuringiensis*, which had a larger core proteome than the random sets, but actually had fewer unique proteins than the average number of unique proteins in the random sets. In addition, the *B. thuringiensis* isolates had fewer unique proteins than seven of the 25 corresponding random sets.

Further examination of Tables 5.3 and 5.4 showed that all species satisfied the criterion that species should have proteomes that are similar to each other, as all of the P-values in column P^C were significant at the 5% level. However, the same cannot be said of the unique proteins criterion. Several species, in addition to those already mentioned in the previous paragraph, had either a statistically insignificant difference in unique proteins compared to the random sets, or had very few unique proteins compared to most other species. *Brucella abortus*, *B. suis*, *B. mallei*, *C. botulinum*, *Mycobacterium bovis*, *Mycobacterium tuberculosis*, *Y. pestis*, and *Y. pseudotuberculosis* all fell within this category.

Chapter 6
Conclusions and Discussion

This section provides discussion concerning selected aspects of PSI and suggests possibilities for future work. Section 6.1 provides an overview of PSI, reviews the major contributions of this thesis, gives some concluding remarks, and comments on the general applicability of PSI now and in the future. Section 6.2 proposes as future work a comparison between BLAST and other methods for protein database searching. Modifications to PSI that could make it more efficacious for eukaryotic organisms are suggested in Section 6.3. Section 6.4 discusses the utility of PSI in identifying protein-phenotype relationships for proteins involved in different types of biochemical pathways. Section 6.5 contains a comparison of the AUP metric with the proteomic distance metric proposed by Snel et al., and also suggests possibilities for future work concerning the application of protein content comparisons to phylogenetics. The efficacy of PSI when applied to incomplete datasets is discussed in Section 6.6. Section 6.7 describes several ways in which PSI could be used that were not analyzed in detail in this thesis. Finally, Section 6.8 discusses the nature of the author's collaboration with Monique Haakensen, who is currently completing her Ph.D. thesis.

6.1 Conclusion

This thesis describes the foundation, design, and implementation of a program called PSI that facilitates the discovery of proteins that are found in one set of organisms, but not a second set. Two broad applications of PSI were identified and analyzed: determining protein-phenotype relationships, and comparing the protein content in different groups of organisms. In addition to the creation of PSI, this thesis has made several contributions:

- a method for visualizing orthologous relationships among proteins,
- a demonstration that PSI was useful in identifying the protein responsible for the difference in cell shape between *P. pentosaceus* and *Lactobacillus* isolates,
- the identification of several proteins that could be responsible for gatifloxacin resistance in the LAB *P. pentosaceus* and *L. brevis*,
- analytical and empirical methods for identifying appropriate E-value thresholds for PSI,

- a novel metric for measuring differences in protein content in pairs of organisms,
- the largest phylogenetic tree created to date (to the author's knowledge) based on protein content,
- a technique for assessing how well different species are clustered based on protein content, and
- a list of species whose taxonomic classifications may warrant reexamination based on the protein content clustering analysis.

In conclusion, PSI is an extremely useful tool for addressing many comparative genomics questions and for discovering protein-phenotype relationships. Besides those analyzed in this thesis, PSI has a number of additional applications. For instance, it could be used for measuring the general diversity of protein content in different organisms, comparing intra-species with inter-species proteomic diversity, comparing the metabolism and physiology of different genera, characterizing the impact of environment on protein content, and even tracking evolution on a fine-grained scale. A more detailed description of these proposed applications can be found in Section 6.7.

Due to its generality, the number of possible applications for PSI is extremely large, and there likely are many applications for PSI beyond those analyzed or suggested in this thesis. Perhaps the most exciting aspect of PSI, as well as other tools that utilize genomic sequence information, is that its usefulness will only continue to increase as more and more sequence information becomes available. Despite advances in sequencing technology, the number of species whose genomes have been sequenced to date is likely just a tiny fraction of the number of species that currently exist on Earth [74]. Even so, much progress has been made in learning about the genetic and molecular properties of the organisms whose genomes have been sequenced. While the prospect of elucidating the complete genome sequences of a substantial portion of extant species is still a long way off, it is certainly a tantalizing proposition. Using programs like PSI, much profound knowledge could be gained—for instance, the number of proteins comprising the entire protein universe could be estimated; the uniqueness in the protein content of each species could be quantified; the set of proteins that are absolutely necessary for life could be identified; and evolutionary relationships could be delineated more accurately. Thus, tools like PSI will become even more valuable in the future for learning about protein function and for understanding the molecular relationships among different organisms.

6.2 Comparing methods for protein database searching

Section 5.1.1 reported the results of comparing the efficiency of DFS on a graph with that of a disjoint-set data structure for finding orthologous groups of proteins. This operation had to be performed after orthologous relationships between pairs of proteins had been ascertained using RBH. Another aspect of PSI's efficiency that could be examined concerns the BLAST comparisons themselves. Price et al. [75] developed FastBLAST, which attempts to provide faster database searches than BLAST by using known protein families from sources like the Pfam [76, 77] and protein analysis through evolutionary relationships (PANTHER) [78, 79] databases, as well as ad-hoc families that capture homology relationships not found in these databases. More specifically, for a given query sequence, FastBLAST finds the best hits in a database by inspecting only the protein families that the query sequence belongs to, rather than all of the proteins in the database. Since a given protein will likely belong to only a few families, this strategy substantially reduces the number of comparisons that must be performed. FastBLAST may miss some homologues that ordinary BLAST discovers if a particular database protein is not in the same family (in any of the protein family databases) as the query protein. However, the authors of FastBLAST show that such misses are rare, with FastBLAST finding about 98% of the matches that BLAST discovers.

Despite the straightforward concept behind FastBLAST, comparing BLAST with FastBLAST would likely prove to be quite involved. Unlike BLAST, FastBLAST requires a substantial amount of preprocessing to be done before the actual database searches can be performed. Thus, the speed of a database search using BLAST could not be compared directly with a database search using FastBLAST, as the time taken to perform the preprocessing steps that FastBLAST requires would also have to be taken into account. These preprocessing steps are as follows:

1. The input FASTA files must be processed so that the headers conform to the requirements of some of the external programs that FastBLAST uses.

2. A program employing hidden Markov models (HMMs) called FastHMM must be utilized in order to determine the protein family or families to which each database sequence belongs.

3. Alignments of the query sequences against other proteins from the same family must be created.

A fair comparison of BLAST with FastBLAST would have to balance the time that FastBLAST saves doing the actual database searches with the extra time spent performing these preprocessing steps. Note that FastBLAST requires a fourth preprocessing step, which is the creation of indexed databases from multi-FASTA files. However, BLAST also requires this step, so it is not included in the above list.

Another complication is that, due to having a number of fixed costs related to processing the protein families [75], FastBLAST's efficiency decreases with smaller databases. The creators of FastBLAST showed that it exhibited increased efficiency over BLAST when using "nr", the nonredundant GenBank database; however, it would likely be less efficient for situations in which many smaller databases are needed (such as PSI, which uses a separate database for each organism's proteome). Thus, the sizes of the protein databases would also need to be taken into account. Despite these complexities, comparing BLAST with FastBLAST would certainly make worthwhile future work, as it could clarify which problems would be better suited to BLAST, and which would be better suited to FastBLAST.

Yet another method that could be explored for quickly finding homology relationships is CD-HIT [80], which uses a technique called "short word filtering" for quickly clustering similar sequences. Unlike FastBLAST, whose results should be very similar to that of BLAST, the results of using CD-HIT could be quite different. Thus, unlike FastBLAST, which could be compared to BLAST solely in terms of efficiency, an analysis of CD-HIT would require both its speed and its functionality to be compared to those of BLAST.

6.3 PSI and eukaryotic proteomes

In this thesis, PSI was evaluated entirely using prokaryotic proteomes (and more specifically, bacterial proteomes). While the general idea behind PSI should apply equally well to eukaryotes, some modifications may be necessary in order for maximize PSI's efficacy when applied to these more complex organisms. For instance, a new E-value threshold might be appropriate in order to reflect the larger proteome sizes of eukaryotes. In Section 4.1.6, an upper limit on proteome sizes of 10^5 (which is larger than the largest bacterial proteome) was used in order to derive an appropriate E-value threshold. This number would have to be revised when considering eukaryotes, as many eukaryotes have proteomes containing more than 10^5 proteins. The human proteome and the mouse proteome, for instance, each contain around 30000 proteins [81]. Thus, selecting a new E-value threshold based on the size of eukaryotic proteomes would be necessary.

Perhaps the most important issue that would need to be addressed before PSI could be applied to eukaryotes is orthologue detection. Orthologue detection is a fundamental part of PSI, and this procedure is more difficult for eukaryotes than it is for prokaryotes [47]. There are several reasons for this: eukaryotic genomes are much larger; gene duplications are more frequent than in prokaryotes; alternative splicing can occur, in which a single gene can code for more than one protein; eukaryotes are often diploid (have two sets of chromosomes) or polyploid (have more than two sets of chromosomes); and their proteins generally have a more complex architecture, often containing functional domains that are present in many different proteins. To address this added difficulty, future work could involve modifying PSI to use one of the orthologue detection techniques that have been specifically designed to address the problems inherent in predicting orthologues in eukaryotes. As these methods are more sophisticated, this could also result in improved orthologue detection for prokaryotes as well.

6.4 Identifying protein-phenotype relationships in different types of biochemical pathways

Sections 5.2.1 and 5.2.3 presented the results of using PSI for (respectively) finding the protein(s) responsible for giving lactobacilli rod-shaped cells, and for finding protein(s) responsible for gatifloxacin resistance in the two LAB *L. brevis* and *P. pentosaceus.* The rod-like shape of *Lactobacillus* cells appears to be caused by one or both of the MreB-like proteins, while the nature of gatifloxacin resistance in the aforementioned bacteria remains uncertain.

It is interesting to consider how well PSI would work for phenotypes that arise from different, and perhaps more complicated, types of biochemical pathways. There are several possible situations, each of which would differ in the ability of PSI to correctly identify the protein or proteins responsible for the phenotype of interest.

6.4.1 One protein directly causes the phenotype of interest

In the first possible situation, there is only one protein responsible for the phenotype of interest, and the presence of the protein itself (rather than the product of some reaction catalyzed by that protein) causes the phenotype. For instance, suppose there exists a hypothetical protein that causes a bacterium to appear blue under ultraviolet light. PSI should work very well in situations such as this. By identifying proteins that are found in bacteria that appear blue under ultraviolet light, but that are absent from bacteria that do not exhibit this property, PSI should be successful in identifying the correct protein. Note that the difference in cell shape between *Lactobacillus* isolates and *P. pentosaceus* almost fits this category. However, the results given in Section 5.2.1 suggest that the situation is a little more complicated, as two MreB-like proteins were identified as being present in all of the lactobacilli, but not *P. pentosaceus.* In addition, other proteins were identified that could potentially contribute to cell shape, such as sepF. As such, while it is possible that just a single protein is responsible for the difference in cell shape between the lactobacilli and *P. pentosaceus* (as in the hypothetical ultraviolet light example given above), it is also possible that the situation is more complex.

6.4.2 One protein catalyzes the synthesis of a molecule causing the phenotype of interest

Some phenotypes are caused not by proteins, but by other molecules. Suppose that a given phenotype is caused by molecule B, which can only be synthesized from molecule A. Further, this synthesis must be catalyzed by an enzyme denoted $\alpha\beta$. For the sake of this example, assume that A is a ubiquitous cellular metabolite present in all organisms; otherwise, the presence or absence of other proteins may affect the availability of molecule A, making the situation more complicated. All of the organisms that exhibit this phenotype should have protein $\alpha\beta$, while all those that do not should lack this protein; as such, PSI should be successful in this situation, as the presence or absence of $\alpha\beta$ dictates the presence or absence of molecule B, which in turns determines the presence or absence of the phenotype.

6.4.3 Two proteins catalyze the synthesis of a molecule causing the phenotype of interest

Suppose, in contrast to the previous situation, that molecule B can be synthesized from two different ubiquitous cellular metabolites, denoted A_1 and A_2. Also, suppose that the reactions $A_1 \rightarrow B$ and $A_2 \rightarrow B$ are catalyzed by two nonorthologous proteins denoted $\alpha_1\beta$ and $\alpha_2\beta$. PSI may or may not be able to identify these proteins as being responsible for the phenotype induced by molecule B. If all of the organisms that exhibit the phenotype produce B from A_1 using protein $\alpha_1\beta$, then PSI would be able to identify $\alpha_1\beta$ as the protein responsible. Similarly, PSI could identify $\alpha_2\beta$ as the protein responsible if all of the organisms that exhibit the phenotype produce B from A_2. However, if some of the organisms exhibiting the phenotype produce B from A_1, and others produce B from A_2, then PSI would be able to identify neither $\alpha_1\beta$ nor $\alpha_2\beta$. This is because the organisms synthesizing B from A_1 would not contain $\alpha_2\beta$, and the organisms synthesizing B from A_2 would not contain $\alpha_1\beta$, and thus neither protein would be found in all of the organisms exhibiting the phenotype.

Note that this type of problem would occur for any situation in which a phenotype has more than one cause, not just those that involve the catalysis of metabolites. For instance, one protein could cause resistance to a given antibiotic by pumping it out of the cell, whereas another protein could cause resistance by destroying the antibiotic. If some of the organisms in a given comparison are resistant because of the first protein, and some are resistant because of the second, then neither protein would be identified by PSI as causing resistance to the antibiotic, as PSI requires that candidate proteins be present in all organisms exhibiting the phenotype.

6.4.4 The phenotype of interest is part of a long biochemical pathway

The situation would become more complicated if the molecule causing the phenotype was the end product of a longer or more complicated biochemical pathway. For example, suppose that a given organism encodes enzymes $\alpha\beta$, $\beta\gamma$, and $\gamma\delta$, which catalyze the chemical reactions $A \rightarrow B$, $B \rightarrow C$, and $C \rightarrow D$, respectively. Further, suppose that the actual phenotype of interest is caused by molecule D. Ideally, PSI would be able to identify all three proteins as being involved with the creation of this phenotype. In an ideal situation, all of the organisms that do not exhibit the phenotype would contain none of the aforementioned proteins. However, molecules B and C may also have important biochemical roles, and may be synthesized as an end product by one or more of the organisms not exhibiting the phenotype of interest. If this is the case, $\alpha\beta$ and $\beta\gamma$ would not be identified by PSI. However, $\gamma\delta$ would still be identified, as this protein would be expressed by all of the organisms that exhibit the phenotype, and none of the organisms that do not. This result would probably be satisfactory to most users, as $\gamma\delta$ is the protein most directly responsible for the phenotype of interest, since it catalyzes the creation of the molecule that actually produces the phenotype.

6.5 Proteomic distance metrics

As described in Section 4.3.3, Snel et al. [62] used the number of shared proteins between two organisms divided by the number of proteins in the smaller proteome as a metric for determining the proteomic distance between two organisms. In this thesis, an alternative metric, AUP, was proposed, which is calculated by taking the average of the number of proteins in bacterium A that are not in bacterium B, and the number of proteins in bacterium B that are not in bacterium A. This new metric was proposed because it fits well within the scope of "genome subtraction", in which proteins are found that are present in all of the organisms in one set of organisms, but in none of the organisms in a second set. In the case of pairwise comparisons, each set contains just one organism. This metric was used to create a phylogenetic tree of the bacteria listed in Figure 4.2. This phylogenetic tree (see Figure 5.17) appeared to be relatively consistent with current taxonomic classifications, with most isolates of the same species clustering together. However, a more in-depth analysis would be needed to determine how this metric compares with the one proposed by Snel et al.. This section discusses two possible levels at which proteomic distance metrics could be compared—on the level of individual pairwise comparisons (Section 6.5.1), and on the level of phylogenetic trees created using the distance metrics (Section 6.5.2). In addition, Section 6.5.3 discusses finding the correlation between proteomic distance metrics and 16S rRNA gene percent identities for pairs of bacteria.

6.5.1 Comparing proteomic distance metrics at the level of individual pairwise comparisons

The first way in which the proteomic distance metrics can be compared is by looking at pairwise comparisons. For instance, consider the question of whether *Rickettsia typhi* ATCC VR-144 is more closely related to *L. gasseri* ATCC 33323 or to *Burkholderia vietnamiensis* LMG 2248. The number of proteins that are common to both *R. typhi* and *L. gasseri* is 260, while the number common to both *R. typhi* and *B. vietnamiensis* is 437. *R. typhi*, with 837 proteins, has the smaller proteome in each of these comparisons, so the values of Snel et al.'s metric for these two comparisons are

Table 6.1: Comparison of the distance metric of Snel et al. with the AUP metric. See below the table for an explanation of the column headings.

Reference organism	**Comparison organism 1**	$\mathbf{S_1}$	$\mathbf{A_1}$	**Comparison organism 2**	$\mathbf{S_2}$	$\mathbf{A_2}$
R. typhi	*L. gasseri*	0.31	1005.5	*B. vietnamiensis*	0.52	3686
V. cholerae	*Y. pestis*	0.43	2223.5	*B. ambifaria*	0.34	3903.5
L. acidophilus	*S. pyogenes*	0.37	1188.5	*R. etli*	0.26	3403
R. etli	*B. ambifaria*	0.31	4424	*S. pyogenes*	0.27	3398.5

Column heading abbreviations are as follows: S_1, similarity between the reference organism and comparison organism 1 according to Snel et al.'s metric; A_1, similarity between the same organisms according to the AUP metric; S_2, similarity between the reference organism and comparison organism 2 according to Snel et al.'s metric; A_2, similarity between the same organisms according to the AUP metric. Note that for Snel et al.'s metric, larger numbers indicate greater similarity, whereas for the AUP metric, smaller numbers indicate greater similarity. To enable the table to fit on the page, strain names are omitted. The full designation of the organisms in the table, as well as the number of proteins found in the proteome of each, are *Rickettsia typhi* ATCC VR-144 (837 proteins), *L. gasseri* ATCC 33323 (1694 proteins), *Burkholderia vietnamiensis* LMG 22486 (7409 proteins), *Vibrio cholerae* ATCC 39315 (3784 proteins), *Yersinia pestis* 91001 (4013 proteins), *Burkholderia ambifaria* ATCC BAA-244 (6607 proteins), *Lactobacillus acidophilus* NCFM (1859 proteins), *Streptococcus pyogenes* MGAS2096 (1886 proteins), and *Rhizobium etli* ATCC 51251 (5921 proteins).

$260/837 = 0.31$ and $437/837 = 0.52$, respectively.

Does the fact that *R. typhi* shares more proteins with *B. vietnamiensis* imply that *R. typhi* is more closely related to *B. vietnamiensis* than it is to *L. gasseri*? This is unclear; however, a plausible explanation for the disparity in these two numbers is the size of the proteome of *L. gasseri* (1694 proteins) compared to that of *B. vietnamiensis* (7409 proteins)—because the latter bacterium's proteome is so much larger, it has more "opportunities" for containing proteins that are orthologous to proteins in *R. typhi.* In this case, Snel et al.'s metric seems like it is being influenced more by proteome sizes than by actual proteomic similarity. In fact, the extremely large proteome of *B. vietnamiensis* would perhaps suggest that *R. typhi*, whose proteome contains just 837 proteins, is more distant evolutionarily to *B. vietnamiensis* than it is to *L. gasseri.*

In contrast, the AUP metric produces quite different results for the above example. The number of proteins in *R. typhi* that are not in *L. gasseri* is 577, while the number of proteins in *L. gasseri* that are not in *R. typhi* is 1434, giving an average of 1005.5. On the other hand, the number of proteins in *R. typhi* that are not in *B. vietnamiensis* is 400, while the number of proteins in *B. vietnamiensis* that are not in *R. typhi* is 6972, giving an average of 3686. Contrary to the method of Snel et al., this suggests that *R. typhi* is more closely related to *L. gasseri* than it is to *B. vietnamiensis*, a result that seems more consistent with intuition.

Table 6.1 gives three additional examples, and also reiterates the above example. The first row of the table (not including the row containing the column headings) is the same as the example given above, in which Snel et al.'s metric is inconsistent with the AUP metric, with the former metric suggesting that *R. typhi* is more similar to *B. vietnamiensis* than it is to *L. gasseri*, and the latter metric suggesting the opposite. The next two rows of the table show examples where the two metrics agree. The second row shows that the metrics agree that *Vibrio cholerae* ATCC 39315 is more similar to *Y. pestis* 91001 than it is to *Burkholderia ambifaria* ATCC BAA-244, and the third row shows that they agree that *L. acidophilus* NCFM is more similar to *S. pyogenes* MGAS2096 than it is to *R. etli* ATCC 51251. The fourth row of the table gives another example where the two metrics are discordant: in contrast to Snel et al.'s metric, the AUP metric suggests that *R. etli* ATCC 51251 is more similar to *S. pyogenes* MGAS2096 than it is to *B. ambifaria* ATCC BAA-244. In this case, Snel et al.'s method seems more in agreement with intuition, as it seems more likely that *R. etli*, having a fairly large proteome (5921 proteins), would be more similar to another organism with a large proteome (*B. ambifaria*, with 6607 proteins) than one with a small proteome (*S. pyogenes*, with just 1886 proteins).

In summary, Table 6.1 gives two examples where the two metrics agree, one example where the AUP metric seems to give more intuitive results, and one example where Snel et al.'s metric seems to give more intuitive results. Note that this analysis dealt only with whether the results agree with intuition, and a more rigorous analysis would be needed to compare the results of both metrics with, say, 16S rRNA gene percent identity. Overall, additional investigation will be necessary in order to analyze these metrics further and to elucidate the most appropriate method for determining the proteomic similarity (or distance) between two organisms.

6.5.2 Comparing proteomic distance metrics at the level of phylogenetic trees

Section 6.5.1 discussed differences between Snel et al.'s metric and the AUP metric for a few specific pairwise comparisons, and suggested that more work needs to be done in order to establish the strengths and weaknesses of each of these metrics, as well as to explore other metrics that might provide a more accurate measure of the proteomic similarity of two organisms. It would also be interesting to compare the two distance metrics (as well as other possible metrics) on a broader scale, by comparing the phylogenetic trees that result from these metrics. The tree given in Figure 5.17 seems plausible, with most isolates from the same genus clustering together. Future work could involve constructing a phylogenetic tree using the same organisms, but using Snel et al.'s distance metric, and then comparing the two trees. It would also be worthwhile to compare the two trees to a tree created using 16S rRNA gene percent identities.

6.5.3 Comparing proteomic distance metrics with 16S rRNA gene percent identities

As examining changes in the 16S rRNA gene is the standard method for performing phylogenetic analyses, it would be of interest to determine how well different proteomic distance metrics correlate with percent identities between the 16S rRNA genes in pairs of organisms. The thesis of Monique Haakensen (see also Section 6.8) examines the correlation between the AUP metric and 16S rRNA gene percent identity for intra-genus pairwise comparisons. It was found that, for some genera, the correlation between the AUP metric and the percent identity of the 16S rRNA genes was reasonable (values of R^2 between 0.47 and 0.81), whereas other genera had very low R^2 values (close to zero). Further investigation would be necessary in order to elucidate the reasons for the different correlations among the genera; examining the amount of horizontal gene transfer that occurs in different genera would seem like a reasonable starting point. It would also be interesting to find the correlation of Snel et al.'s metric with 16S rRNA gene percent identities.

6.6 PSI and incomplete datasets

Despite the accelerating pace of genome sequencing, only partial sequence information is available for some organisms. This could be the result of researchers sequencing only certain genes or other genomic regions that are of interest to them, rather than an entire genome. In addition, some organisms are studied primarily using complementary DNA (cDNA) or expressed sequence tag (EST) libraries, which do not necessarily represent all of the genes in a genome; furthermore, ESTs do not usually encode entire proteins.

In general, PSI would be expected to have limited usefulness for organisms lacking a complete proteome. The goal of PSI is to identify proteins that are present in one set of organisms, but not a second set. If one or more of these organisms have incomplete proteomes, then the absence of a specific protein in a particular organism could be due to incomplete sequence information, or because its genome actually does not encode that protein. PSI could not distinguish between these two possibilities, potentially leading to erroneous results. However, as more genomes become sequenced, partial sequence information should become less of an issue.

6.7 Other comparative genomics applications

Besides the applications already examined or discussed in this thesis, the versatile nature of PSI means that it should be useful for addressing many other issues concerning the protein content in groups of organisms. A list of some of these applications is given in this section. This list should not, however, be considered complete; there are likely many other potential applications of PSI not mentioned in this thesis.

6.7.1 Measuring diversity of protein content

PSI could be used to determine, for each protein in a given organism, whether it is unique to that organism, unique to its species, unique to its genus, or non-unique (found in other genera). This would constitute a general characterization of the amount of protein uniqueness and redundancy in the universe of bacterial proteins. PSI could also be used to find the total number of distinct proteins present in a given species or genus, which would provide additional insight into the diversity of protein content in different groups of organisms.

6.7.2 Measuring intra-species versus inter-species protein diversity

PSI would be useful in determining whether, within a given genus, the number of proteins found in organism A but not in organism B is always smaller if A and B are from the same species than if they are from different species. This represents another possible method for evaluating the quality of phylogenetic classifications—if organism A from species X has more proteins not found in organism B (from species X) than organism C (from species Y), then it could indicate the need to revisit the taxonomic classifications of these organisms. Situations like this could also be a result of differences in the rate of changes to protein content compared to the rate of changes to the 16S rRNA gene. As such, this could also provide insight into the relative rates of protein content evolution versus evolution of the 16S rRNA gene.

6.7.3 Comparing the protein content of different genera

PSI could be used to answer the question, "What kinds of proteins are found in all of the isolates in genus A, but in none of the isolates of genus B?" Examining such differences in protein content would allow the metabolism and physiology of different genera to be compared, and could also be useful for evaluating the impact of environment on protein content. For instance, a possible study could involve comparing the protein content in pairs of genera that exist in similar environments, as well as in pairs that exist in disparate environments. One might expect that pairs of genera inhabiting similar environments would have more similar protein content than those that inhabit disparate environments. By comparing protein content differences with evolutionary differences, the impact of environment on protein content could be characterized.

6.7.4 Tracking evolution

Given recent technological developments in sequencing technologies, it is reasonable to assume that, at some point in the future, sequencing a bacterium will require only a nominal investment of time and money. The following methodology involving PSI would enable tracking of the evolution of a particular species. First, a single bacterium from the species would be sequenced. Then, a sample of its descendants would periodically be sequenced (say, every few weeks or months). PSI would be used to compare the protein content in these organisms, which would give a fine-grained look at what proteins were acquired or lost over these time periods. If the organisms were under some type of environmental stress, it would allow monitoring of the evolutionary response to the stress at the protein level. Another interesting idea would be to subject some of the bacteria to a change of conditions, such as different growth media, antibiotics, and so on, and then to track the proteins gained or lost.

6.8 Nature of collaboration

The work described in this thesis has been applied in the Ph.D. thesis of Monique Haakensen. Specifically, I used the technique described in Section 4.3.3 to create a phylogenetic tree of the 15 sequenced *Lactobacillus* isolates and *P. pentosaceus*. This tree was relatively consistent with those made using the 16S rRNA gene, and supported her contention that these organisms merit a new taxonomic classification. Regarding the analysis described in Section 6.5.3, PSI was used to generate the "unique proteins" data, and Ms. Haakensen determined the 16S rRNA gene percent identities and performed the actual correlation analysis. Her thesis also discusses the application of PSI to finding protein-phenotype relationships and to finding core proteomes. Some of the figures created for this thesis were also included in Ms. Haakensen's thesis.

I am responsible for all of the figures and writing in this thesis; however, Ms. Haakensen was extremely helpful in suggesting ideas and assisting with biological interpretations. Specifically, she suggested the cell shape phenotype and the gatifloxacin resistance phenotype as test cases for using PSI for identifying protein-phenotype relationships, and also assisted with making graphical representations of phylogenetic trees.

References

[1] D K Apps, B B Cohen, and C M Steel. *Biochemistry*. Bailliere Tindall, 5th edition, 1992.

[2] F Sanger and A R Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol*, 94(3):441–448, 1975. ISSN 0022-2836 (Print).

[3] F Sanger, G M Air, B G Barrell, N L Brown, A R Coulson, C A Fiddes, C A Hutchison, P M Slocombe, and M Smith. Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature*, 265 (5596):687–695, 1977. ISSN 0028-0836 (Print).

[4] F Sanger, S Nicklen, and A R Coulson. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12):5463–5467, 1977. ISSN 0027-8424 (Print).

[5] F Sanger, A R Coulson, G F Hong, D F Hill, and G B Petersen. Nucleotide sequence of bacteriophage λ DNA. *J Mol Biol*, 162(4):729–773, 1982. ISSN 0022-2836 (Print).

[6] S J Goebel, G P Johnson, M E Perkus, S W Davis, J P Winslow, and E Paoletti. The complete DNA sequence of vaccinia virus. *Virology*, 179(1):247–266, 1990. ISSN 0042-6822 (Print).

[7] R F Massung, L I Liu, J Qi, J C Knight, T E Yuran, A R Kerlavage, J M Parsons, J C Venter, and J J Esposito. Analysis of the complete genome of smallpox variola major virus strain Bangladesh-1975. *Virology*, 201(2):215–240, 1994. ISSN 0042-6822 (Print). doi: 10.1006/viro.1994.1288.

[8] R Staden. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res*, 6(7):2601–2610, 1979. ISSN 0305-1048 (Print).

[9] H Kasai, S Isono, M Kitakawa, J Mineno, H Akiyama, D M Kurnit, D E Berg, and K Isono. Efficient large-scale sequencing of the *Escherichia coli* genome: implementation of a transposon- and PCR-based strategy for the analysis of ordered λ phage clones. *Nucleic Acids Res*, 20(24): 6509–6515, 1992. ISSN 0305-1048 (Print).

[10] V Burland, D L Daniels, G Plunkett III, and F R Blattner. Genome sequencing on both strands: the Janus strategy. *Nucleic Acids Res*, 21(15):3385–3390, 1993. ISSN 0305-1048 (Print).

[11] J C Roach, C Boysen, K Wang, and L Hood. Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics*, 26(2):345–353, 1995. ISSN 0888-7543 (Print).

[12] R D Fleischmann, M D Adams, O White, R A Clayton, E F Kirkness, A R Kerlavage, C J Bult, J F Tomb, B A Dougherty, and J M Merrick. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269(5223):496–512, 1995. ISSN 0036-8075 (Print).

[13] A Goffeau, B G Barrell, H Bussey, R W Davis, B Dujon, H Feldmann, F Galibert, J D Hoheisel, C Jacq, M Johnston, E J Louis, H W Mewes, Y Murakami, P Philippsen, H Tettelin, and S G Oliver. Life with 6000 genes. *Science*, 274(5287):546, 563–7, 1996. ISSN 0036-8075 (Print).

[14] M D Adams *et al.* The genome sequence of *Drosophila melanogaster*. *Science*, 287(5461): 2185–2195, 2000. ISSN 0036-8075 (Print).

[15] The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814):796–815, 2000. ISSN 0028-0836 (Print). doi: 10.1038/35048692.

[16] E S Lander *et al.* Initial sequencing and analysis of the human genome. *Nature*, 409(6822): 860–921, 2001. ISSN 0028-0836 (Print). doi: 10.1038/35057062.

[17] J C Venter et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001. ISSN 0036-8075 (Print). doi: 10.1126/science.1058040.

[18] P Kersey, L Bower, L Morris, A Horne, R Petryszak, C Kanz, A Kanapin, U Das, K Michoud, I Phan, A Gattiker, T Kulikova, N Faruque, K Duggan, P Mclaren, B Reimholz, L Duret, S Penel, I Reuter, and R Apweiler. Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res*, 33(Database issue):D297–302, 2005. ISSN 1362-4962 (Electronic). doi: 10.1093/nar/gki039.

[19] European Bioinformatics Institute. Genomes pages - bacteria (http://www.ebi.ac.uk/genomes/bacteria.html), 2009.

[20] European Bioinformatics Institute. Genomes pages - Archaea (http://www.ebi.ac.uk/genomes/archaea.html).

[21] European Bioinformatics Institute. Genomes pages - virus (http://www.ebi.ac.uk/genomes/virus.html), 2009.

[22] European Bioinformatics Institute. Genomes pages - eukaryota (http://www.ebi.ac.uk/genomes/eukaryota.html), 2009.

[23] R L Tatusov, E V Koonin, and D J Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, 1997. ISSN 0036-8075 (Print).

[24] S B Needleman and C D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453, 1970. ISSN 0022-2836 (Print).

[25] T F Smith and M S Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197, 1981. ISSN 0022-2836 (Print).

[26] D J Lipman and W R Pearson. Rapid and sensitive protein similarity searches. *Science*, 227 (4693):1435–1441, 1985. ISSN 0036-8075 (Print).

[27] W R Pearson and D J Lipman. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8):2444–2448, 1988. ISSN 0027-8424 (Print).

[28] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, 1990. ISSN 0022-2836 (Print). doi: 10.1006/jmbi.1990.9999.

[29] S F Altschul, T L Madden, A A Schaffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, 1997. ISSN 0305-1048 (Print).

[30] GraphViz (http://www.graphviz.org/), 2008.

[31] K S Makarova and E V Koonin. Evolutionary genomics of lactic acid bacteria. *J Bacteriol*, 189(4):1199–1208, 2007. ISSN 0021-9193 (Print). doi: 10.1128/JB.01351-06.

[32] R L Tatusov, M Y Galperin, D A Natale, and E V Koonin. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*, 28(1):33–36, 2000. ISSN 0305-1048 (Print).

[33] R L Tatusov, D A Natale, I V Garkavtsev, T A Tatusova, U T Shankavaram, B S Rao, B Kiryutin, M Y Galperin, N D Fedorova, and E V Koonin. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*, 29(1):22–28, 2001. ISSN 1362-4962 (Electronic).

[34] R L Tatusov, N D Fedorova, J D Jackson, A R Jacobs, B Kiryutin, E V Koonin, D M Krylov, R Mazumder, S L Mekhedov, A N Nikolskaya, B S Rao, S Smirnov, A V Sverdlov, S Vasudevan, Y I Wolf, J J Yin, and D A Natale. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41, 2003. ISSN 1471-2105 (Electronic). doi: 10.1186/1471-2105-4-41.

[35] K Makarova, A Slesarev, Y Wolf, A Sorokin, B Mirkin, E Koonin, A Pavlov, N Pavlova, V Karamychev, N Polouchine, V Shakhova, I Grigoriev, Y Lou, D Rohksar, S Lucas, K Huang, D M Goodstein, T Hawkins, V Plengvidhya, D Welker, J Hughes, Y Goh, A Benson, K Baldwin, J-H Lee, I Diaz-Muniz, B Dosti, V Smeianov, W Wechter, R Barabote, G Lorca, E Altermann, R Barrangou, B Ganesan, Y Xie, H Rawsthorne, D Tamir, C Parker, F Breidt, J Broadbent, R Hutkins, D O'Sullivan, J Steele, G Unlu, M Saier, T Klaenhammer, P Richardson, S Kozyavkin, B Weimer, and D Mills. Comparative genomics of the lactic acid bacteria. *Proc Natl Acad Sci U S A*, 103(42):15611–15616, 2006. ISSN 0027-8424 (Print). doi: 10.1073/pnas.0607117103.

[36] K S Makarova, A V Sorokin, P S Novichkov, Y I Wolf, and E V Koonin. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol Direct*, 2:33, 2007. ISSN 1745-6150 (Electronic). doi: 10.1186/1745-6150-2-33.

[37] National Center for Biotechnology Information. COGs - Clusters of Orthologous Groups (http://www.ncbi.nlm.nih.gov/COG/), 2008.

[38] V M Markowitz, F Korzeniewski, K Palaniappan, E Szeto, G Werner, A Padki, X Zhao, I Dubchak, P Hugenholtz, I Anderson, As Lykidis, K Mavromatis, N Ivanova, and N C Kyrpides. The integrated microbial genomes (IMG) system. *Nucleic Acids Res*, 34(Database issue): D344–8, 2006. ISSN 1362-4962 (Electronic). doi: 10.1093/nar/gkj024.

[39] V M Markowitz, E Szeto, K Palaniappan, Y Grechkin, K Chu, I A Chen, I Dubchak, I Anderson, A Lykidis, K Mavromatis, N N Ivanova, and N C Kyrpides. The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Res*, 36(Database issue):D528–33, 2008. ISSN 1362-4962 (Electronic). doi: 10.1093/nar/gkm846.

[40] U.S. Department of Energy Joint Genome Institute. Integrated microbial genomes (http://img.jgi.doe.gov), 2008.

[41] D L Fulton, Y Y Li, M R Laird, B G Horsman, F M Roche, and F S Brinkman. Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics*, 7:270, 2006. ISSN 1471-2105 (Electronic). doi: 10.1186/1471-2105-7-270.

[42] J C Chiu, E K Lee, M G Egan, I N Sarkar, G M Coruzzi, and R DeSalle. OrthologID: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics*, 22 (6):699–707, 2006. ISSN 1367-4803 (Print). doi: 10.1093/bioinformatics/btk040.

[43] C M Zmasek and S R Eddy. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, 3:14, 2002. ISSN 1471-2105 (Electronic).

[44] C E V Storm and E L L Sonnhammer. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, 18(1):92–99, 2002. ISSN 1367-4803 (Print).

[45] M Remm, C E Storm, and E L Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, 314(5):1041–1052, 2001. ISSN 0022-2836 (Print). doi: 10.1006/jmbi.2000.5197.

[46] K P O'Brien, M Remm, and E L L Sonnhammer. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res*, 33(Database issue):D476–80, 2005. ISSN 1362-4962 (Electronic). doi: 10.1093/nar/gki107.

[47] F Chen, A J Mackey, J K Vermunt, and D S Roos. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, 2(4):e383, 2007. ISSN 1932-6203 (Electronic). doi: 10.1371/journal.pone.0000383.

[48] C R Woese. Bacterial evolution. *Microbiol Rev*, 51(2):221–271, 1987. ISSN 0146-0749 (Print).

[49] M C Maiden, J A Bygraves, E Feil, G Morelli, J E Russell, R Urwin, Q Zhang, J Zhou, K Zurth, D A Caugant, I M Feavers, M Achtman, and B G Spratt. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A*, 95(6):3140–5, 1998.

[50] M C Enright and B G Spratt. Multilocus sequence typing. *Trends Microbiol*, 7(12):482–7, 1999.

[51] I Letunic and P Bork. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1):127–8, 2007. doi: 10.1093/bioinformatics/btl529.

[52] S Karlin and C Burge. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet*, 11(7):283–290, 1995. ISSN 0168-9525 (Print).

[53] M W J van Passel, E E Kuramae, A C M Luyf, A Bart, and T Boekhout. The reach of the genome signature in prokaryotes. *BMC Evol Biol*, 6:84, 2006. ISSN 1471-2148 (Electronic). doi: 10.1186/1471-2148-6-84.

[54] S Karlin, J Mrazek, and A M Campbell. Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol*, 179(12):3899–3913, 1997. ISSN 0021-9193 (Print).

[55] P Vandamme, B Pot, M Gillis, P de Vos, K Kersters, and J Swings. Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiol Rev*, 60(2):407–438, 1996. ISSN 0146-0749 (Print).

[56] F Wright. The 'effective number of codons' used in a gene. *Gene*, 87(1):23–29, 1990. ISSN 0378-1119 (Print).

[57] T Coenye and P Vandamme. Extracting phylogenetic information from whole-genome sequencing projects: the lactic acid bacteria as a test case. *Microbiology*, 149(12):3507–3517, 2003. ISSN 1350-0872 (Print).

[58] M Suyama and P Bork. Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends Genet*, 17(1):10–13, 2001. ISSN 0168-9525 (Print).

[59] J Qi, B Wang, and B-I Hao. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J Mol Evol*, 58(1):1–11, 2004. ISSN 0022-2844 (Print). doi: 10.1007/s00239-003-2493-7.

[60] J Qi, H Luo, and B Hao. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Res*, 32(Web Server issue):W45–7, 2004. doi: 10.1093/nar/gkh362.

[61] T Coenye, D Gevers, Y Van de Peer, P Vandamme, and J Swings. Towards a prokaryotic genomic taxonomy. *FEMS Microbiol Rev*, 29(2):147–167, 2005. ISSN 0168-6445 (Print). doi: 10.1016/j.femsre.2004.11.004.

[62] B Snel, P Bork, and M A Huynen. Genome phylogeny based on gene content. *Nat Genet*, 21 (1):108–110, 1999. ISSN 1061-4036 (Print). doi: 10.1038/5052.

[63] C H House and S T Fitz-Gibbon. Using homolog groups to create a whole-genomic tree of free-living organisms: an update. *J Mol Evol*, 54(4):539–547, 2002. ISSN 0022-2844 (Print). doi: 10.1007/s00239-001-0054-5.

[64] S R Henz, D H Huson, A F Auch, K Nieselt-Struwe, and S C Schuster. Whole-genome prokaryotic phylogeny. *Bioinformatics*, 21(10):2329–2335, 2005. ISSN 1367-4803 (Print). doi: 10.1093/bioinformatics/bth324.

[65] T H Cormen, C E Leiserson, R L Rivest, and C Stein. *Introduction to algorithms*. The MIT Press, 2nd edition, 2001.

[66] M A Huynen, Y Diaz-Lazcoz, and P Bork. Differential genome display. *Trends Genet*, 13(10): 389–390, 1997. ISSN 0168-9525 (Print).

[67] L Li, C J Stoeckert, Jr, and D S Roos. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*, 13(9):2178–89, 2003. doi: 10.1101/gr.1224503.

[68] J Blom, SP Albaum, D Doppmeier, A Puhler, FJ Vorholter, M Zakrzewski, and A Goesmann. EDGAR: A software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics*, 10(1):154, 2009. ISSN 1471-2105 (Electronic). doi: 10.1186/1471-2105-10-154.

[69] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29, 2000. ISSN 1061-4036 (Print). doi: 10.1038/75556.

[70] E Boutet, D Lieberherr, M Tognolli, M Schneider, and A Bairoch. UniProtKB/Swiss-Prot. *Methods Mol Biol*, 406:89–112, 2007. ISSN 1064-3745 (Print).

[71] National Center for Biotechnology Information. The statistics of sequence similarity scores (http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html).

[72] L J Jones, R Carballido-Lopez, and J Errington. Control of cell shape in bacteria: helical, actin-like filaments in Bacillus subtilis. *Cell*, 104(6):913–922, 2001. ISSN 0092-8674 (Print).

[73] P M Hawkey. Mechanisms of quinolone action and microbial response. *J Antimicrob Chemother*, 51 Suppl 1:29–35, 2003. ISSN 0305-7453 (Print). doi: 10.1093/jac/dkg207.

[74] R M May. How many species are there on Earth? *Science*, 241(4872):1441–1449, 1988. doi: 10.1126/science.241.4872.1441.

[75] M N Price, P S Dehal, and A P Arkin. FastBLAST: homology relationships for millions of proteins. *PLoS ONE*, 3(10):e3589, 2008. ISSN 1932-6203 (Electronic). doi: 10.1371/journal.pone.0003589.

[76] E L Sonnhammer, S R Eddy, and R Durbin. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28(3):405–420, 1997. ISSN 0887-3585 (Print).

[77] R D Finn, J Tate, J Mistry, P C Coggill, S J Sammut, H-R Hotz, G Ceric, K Forslund, S R Eddy, E L L Sonnhammer, and A Bateman. The Pfam protein families database. *Nucleic Acids Res*, 36(Database issue):D281–8, 2008. ISSN 1362-4962 (Electronic). doi: 10.1093/nar/gkm960.

[78] P D Thomas, M J Campbell, A Kejariwal, H Mi, B Karlak, R Daverman, K Diemer, A Muruganujan, and A Narechania. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res*, 13(9):2129–2141, 2003. ISSN 1088-9051 (Print). doi: 10.1101/gr.772403.

[79] H Mi, B Lazareva-Ulitsky, R Loo, A Kejariwal, J Vandergriff, S Rabkin, N Guo, A Muruganujan, O Doremieux, M J Campbell, H Kitano, and P D Thomas. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res*, 33(Database issue): D284–8, 2005. ISSN 1362-4962 (Electronic). doi: 10.1093/nar/gki078.

[80] W Li and A Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006. ISSN 1367-4803 (Print). doi: 10.1093/bioinformatics/btl158.

[81] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002. ISSN 0028-0836 (Print). doi: 10.1038/nature01262.

Appendix A

Complete list of organisms used

These tables list the isolates used for some of the analyses described in Sections 4.2 and 4.3. Some strain designations have been removed or shortened to save space. For instance, the full description of the bacterium listed in Table A.3 as "*Burkholderia thailandensis* E264 / ATCC 700388" is actually "*B. thailandensis* (strain E264 / ATCC 700388 / DSM 13276 / CIP 106301)". The name of each organism is accompanied by its taxonomic ID, the number of proteins in its proteome, and its genome size.

Table A.1: Complete list of *Bacillus* isolates used.

TaxID	Isolate	Proteins (#)	Genome size (bp)
326423	*B. amyloliquefaciens* FZB42	3692	3,918,589
261594	*B. anthracis* Ames ancestor	5590	5,227,419
198094	*B. anthracis* Ames, isolate Porton	5313	5,227,293
260799	*B. anthracis* Sterne	5288	5,228,663
222523	*B. cereus* ATCC 10987	5821	5,224,283
226900	*B. cereus* ATCC 14579 / DSM 31	5240	5,411,809
288681	*B. cereus* ZK / E33L	5638	5,300,915
315749	*B. cereus* subsp. cytotoxis, strain NVH 391-98	3840	4,087,024
66692	*B. clausii* KSM-K16	4082	4,303,871
272558	*B. halodurans* C-125 / ATCC BAA-125	4006	4,202,352
279010	*B. licheniformis* DSM 13 / ATCC 14580	4162	4,222,597
315750	*B. pumilus* SAFR-032	3675	3,704,465
224308	*B. subtilis* 168	4112	4,215,606
412694	*B. thuringiensis* Al Hakam	4792	5,257,091
281309	*B. thuringiensis* konkukian, strain 97-27	5169	5,237,682
315730	*B. weihenstephanensis* KBAB4	5650	5,262,775

Table A.2: Complete list of *Brucella* isolates used.

TaxID	Isolate	Proteins (#)	Genome size (bp)
262698	*B. abortus* biovar 1, strain 9-941	3077	3,286,445
359391	*B. abortus* 2308	3022	3,278,307
430066	*B. abortus* S19	2993	3,283,936
483179	*B. canis* ATCC 23365 / NCTC 10854	3238	3,312,769
224914	*B. melitensis* NCTC 10094 / ATCC 23456 / 16M	3178	3,294,931
444178	*B. ovis* ATCC 25840 / 63/290 / NCTC 10512	2820	3,275,590
204722	*B. suis* biovar 1, strain 1330	3256	3,315,175
470137	*B. suis* ATCC 23445 / NCTC 10510	3214	3,324,607

Table A.3: Complete list of *Burkholderia* isolates used.

TaxID	Isolate	Proteins (#)	Genome size (bp)
339670	*B. ambifaria* AMMD / ATCC BAA-244	6607	7,484,986
398577	*B. ambifaria* MC40-6	6690	7,340,944
331271	*B. cenocepacia* AU 1054	6450	7,279,116
331272	*B. cenocepacia* HI2424	6898	7,537,983
406425	*B. cenocepacia* MC0-3	6986	7,971,389
216591	*B. cepacia* J2315 / LMG 16656	6993	7,963,121
243160	*B. mallei* ATCC 23344	4797	5,835,527
412022	*B. mallei* NCTC 10229	5309	5,742,303
320389	*B. mallei* NCTC 10247	5619	5,848,380
320388	*B. mallei* SAVP1	4981	5,232,401
391038	*B. phymatum* DSM 17167 / STM815	7461	6,176,561
398527	*B. phytofirmans* DSM 17436 / PsJN	7197	8,093,536
357348	*B. pseudomallei* 1106a	7138	7,089,249
320372	*B. pseudomallei* 1710b	6329	7,308,054
320373	*B. pseudomallei* 668	7215	7,040,403
272560	*B. pseudomallei* K96243	5717	7,247,547
271848	*B. thailandensis* E264 / ATCC 700388	5561	6,723,972
269482	*B. vietnamiensis* R1808 / G4 / LMG 22486	7409	7,305,580
266265	*B. xenovorans* LB400	8591	9,731,138

Table A.4: Complete list of *Clostridium* isolates used.

TaxID	Isolate	Proteins (#)	Genome size (bp)
272562	*C. acetobutylicum* DSM 792 / JCM 1419	3847	3,940,880
290402	*C. beijerinckii* ATCC 51743 / NCIMB 8052	5003	6,000,632
441770	*C. botulinum* ATCC 19397 / Type A	3547	3,863,450
508767	*C. botulinum* Alaska E43 / type E3	3255	3,659,644
508765	*C. botulinum* Eklund 17B / type B	3525	3,800,327
441771	*C. botulinum* ATCC 3502, substrain Los Alamos	3401	3,760,560
413999	*C. botulinum* ATCC 3502, substrain Sanger	3590	3,886,916
441772	*C. botulinum* Langeland / NCTC 10281 / Type F	3657	3,995,387
498214	*C. botulinum* Loch Maree / Type A3	3982	3,992,906
498213	*C. botulinum* Okra / Type B1	3850	3,958,233
272563	*C. difficile* 630	3712	4,290,252
431943	*C. kluyveri* ATCC 8527 / DSM 555	3828	3,964,618
386415	*C. novyi* NT	2305	2,547,720
195102	*C. perfringens* 13 / Type A	2721	3,031,430
195103	*C. perfringens* ATCC 13124 / NCTC 8237	2873	3,256,683
289380	*C. perfringens* SM101 / Type A	2568	2,897,393
357809	*C. phytofermentans* ATCC 700394	3891	4,847,594
212717	*C. tetani* Massachusetts / E88	2414	2,799,251
203119	*C. thermocellum* ATCC 27405 / DSM 1237	3102	3,843,301

Table A.5: Complete list of *Lactobacillus* isolates used.

TaxID	Isolate	Proteins (#)	Genome size (bp)
272621	*L. acidophilus* NCFM	1859	1,993,560
387344	*L. brevis* ATCC 367 / JCM 1170	2201	2,291,220
321967	*L. casei* ATCC 334	2708	2,895,264
543734	*L. casei* BL23	2999	3,079,196
390333	*L. delbrueckii* ATCC 11842	1519	1,864,998
321956	*L. delbrueckii* ATCC BAA-365	1682	1,856,951
334390	*L. fermentum* IFO 3956 / LMG 18251	1818	2,098,685
324831	*L. gasseri* ATCC 33323 / DSM 20243	1694	1,894,360
405566	*L. helveticus* DPC 4571	1580	2,080,931
257314	*L. johnsonii* NCC 533	1809	1,992,676
220668	*L. plantarum* WCFS1 / ATCC BAA-793	3051	3,308,274
349123	*L. reuteri* 100-23	1972	2,174,299
299033	*L. reuteri* F275	1939	1,999,618
314315	*L. sakei* subsp. sakei, strain 23K	1872	1,884,661
362948	*L. salivarius* subsp. salivarius, strain UCC118	1998	1,827,111

Table A.6: Complete list of *Mycobacterium* isolates used.

TaxID	Isolate	Proteins (#)	Genome size (bp)
36809	*M. abscessus* ATCC 19977 / DSM 44196	4939	5,067,172
243243	*M. avium* 104	5040	5,475,491
233413	*M. bovis* AF2122/97 / ATCC BAA-935	3911	4,345,492
410289	*M. bovis* BCG / Pasteur 1173P2	3891	4,374,522
350054	*M. gilvum* ATCC 700033 / PYR-GCK	5499	5,619,607
272631	*M. leprae* TN	1603	3,268,203
216594	*M. marinum* ATCC BAA-535 / M	5418	6,636,827
262316	*M. paratuberculosis* ATCC BAA-968 / K-10	4316	4,829,781
246196	*M. smegmatis* ATCC 700084 / mc(2)155)	6597	6,988,209
419947	*M. tuberculosis* ATCC 25177 / H37Ra	3990	6,988,209
83332	*M. tuberculosis* ATCC 25618 / H37Rv	3949	6,988,209
83331	*M. tuberculosis* Oshkosh / CDC 1551	4196	4,403,837
362242	*M. ulcerans* Agy99	4206	5,631,606
350058	*M. vanbaalenii* DSM 7251 / PYR-1	5902	6,491,865

Table A.7: Complete list of *Neisseria* isolates used.

TaxID	Isolate	Proteins (#)	Genome size (bp)
242231	*N. gonorrhoeae* ATCC 700825 / FA 1090	1963	2,153,922
521006	*N. gonorrhoeae* NCCP11945	2595	2,232,025
272831	*N. meningitidis* serogroup C, strain ATCC 700532	1865	2,194,961
374833	*N. meningitidis* serogroup C, strain 053442	1998	2,153,416
122587	*N. meningitidis* serogroup A, strain Z2491	1887	2,184,406
122586	*N. meningitidis* serogroup B, strain MC58	2001	2,272,360

Table A.8: Complete list of *Pseudomonas* isolates used.

TaxID	Isolate	Proteins (#)	Genome size (bp)
208964	*P. aeruginosa* LMG 12228 / ATCC 15692	5558	6,264,404
381754	*P. aeruginosa* PA7	6246	6,588,339
208963	*P. aeruginosa* UCBPP-PA14	5886	6,537,648
384676	*P. entomophila* L48	5126	5,888,780
220664	*P. fluorescens* Pf-5 / ATCC BAA-477	6137	7,074,893
205922	*P. fluorescens* PfO-1	5728	6,438,405
399739	*P. mendocina* ymp	4563	5,072,807
351746	*P. putida* F1 / ATCC 700007	5245	5,959,964
76869	*P. putida* GB-1	5396	6,078,430
160488	*P. putida* KT2440	5313	6,181,863
390235	*P. putida* W619	5179	5,774,330
379731	*P. stutzeri* A1501	4093	4,567,418
264730	*P. syringae* 1448A	5044	5,928,787
205918	*P. syringae* pathovar syringae, strain B728a	5071	6,093,698
223283	*P. syringae* tomato, strain DC3000	5424	6,397,126

Table A.9: Complete list of *Rhizobium* isolates used.

TaxID	Isolate	Proteins (#)	Genome size (bp)
347834	*R. etli* CFN 42 / ATCC 51251	5921	4,381,608
491916	*R. etli* CIAT 652	6050	4,513,324
395492	*R. leguminosarum* bv. trifolii WSM2304	4320	4,537,948
216596	*R. leguminosarum* bv. viciae, strain 3841	7109	5,057,142
266835	*R. loti* MAFF303099	7255	7,036,071
266834	*R. meliloti* 1021	6168	3,654,135

Table A.10: Complete list of *Rickettsia* isolates used.

TaxID	Isolate	Proteins (#)	Genome size (bp)
293614	*R. akari* Hartford	1257	1,231,060
391896	*R. bellii* OSU 85-389	1443	1,528,980
336407	*R. bellii* RML369-C	1400	1,522,076
293613	*R. canadensis* McKiel	1091	1,159,772
272944	*R. conorii* ATCC VR-613 / Malish 7	1372	1,268,755
315456	*R. felis* ATCC VR-1525 / URRWXCal2	1428	1,485,148
416276	*R. massiliae* Mtu5	969	1,360,898
272947	*R. prowazekii* Madrid E	834	1,111,523
452659	*R. rickettsii* Iowa	1384	1,268,175
392021	*R. rickettsii* Sheila Smith	1345	1,257,710
257363	*R. typhi* Wilmington / ATCC VR-144	837	1,111,496

Table A.11: Complete list of *Shigella* isolates used.

TaxID	Isolate	Proteins (#)	Genome size (bp)
344609	*S. boydii* serovar 18, strain CDC 3083-94	4140	4,615,997
300268	*S. boydii* serovar 4, strain Sb227	3937	4,519,823
300267	*S. dysenteriae* serovar 1, strain Sd97 / Sd197	3890	4,369,232
198215	*S. flexneri* serovar 2a, strain ATCC 700930	3786	4,599,354
198214	*S. flexneri* serovar 2a, strain 301	4102	4,607,203
373384	*S. flexneri* serovar 5b, strain 8401	3867	4,574,284
300269	*S. sonnei* Ss046	4053	4,825,265

Table A.12: Complete list of *Staphylococcus* isolates used.

TaxID	Isolate	Proteins (#)	Genome size (bp)
93062	*S. aureus* COL	2679	2,809,422
359787	*S. aureus* JH1	2761	2,906,507
359786	*S. aureus* JH9	2708	2,906,700
282458	*S. aureus* MRSA252	2639	2,902,619
282459	*S. aureus* MSSA476	2602	2,799,802
196620	*S. aureus* MW2	2660	2,820,462
418127	*S. aureus* Mu3 / ATCC 700698	2684	2,880,168
158878	*S. aureus* Mu50 / ATCC 700699	2714	2,878,529
158879	*S. aureus* N315	2580	2,814,816
93061	*S. aureus* NCTC 8325	2890	2,821,361
426430	*S. aureus* Newman	2578	2,878,897
451516	*S. aureus* USA300 / TCH1516	2688	2,872,915
451515	*S. aureus* USA300	2607	2,872,769
273036	*S. aureus* bovine RF122 / ET3-1 / RF122	2513	2,742,531
176280	*S. epidermidis* ATCC 12228	2461	2,499,279
176279	*S. epidermidis* ATCC 35984 / RP62A	2492	2,616,530
279808	*S. haemolyticus* JCSC1435	2640	2,685,015
342451	*S. saprophyticus* ATCC 15305	2404	2,516,575

Table A.13: Complete list of *Streptococcus* isolates used.

TaxID	Isolate	Proteins (#)	Genome size (bp)
211110	*S. agalactiae* serovar III, strain NEM316	1999	2,211,485
205921	*S. agalactiae* serovar Ia, strain ATCC 27591	1983	2,127,839
208435	*S. agalactiae* serovar V, strain ATCC BAA-611	2105	2,160,267
552526	*S. equi* MGCS10565	1861	2,024,171
467705	*S. gordonii* ATCC 35105 / CH1	2050	2,196,662
210007	*S. mutans* serovar c, strain ATCC 700610	1951	2,030,921
512566	*S. pneumoniae* serovar 19F, strain G54	2106	2,078,953
373153	*S. pneumoniae* serovar 2, strain NCTC 7466	1918	2,046,115
171101	*S. pneumoniae* ATCC BAA-255 / R6	2030	2,038,615
516950	*S. pneumoniae* CGSP14	2193	2,209,198
487214	*S. pneumoniae* Hungary19A-6	2152	2,245,615
170187	*S. pneumoniae* TIGR4 / ATCC BAA-334	2109	2,160,842
370553	*S. pyogenes* serovar M12, strain MGAS2096	1886	1,860,355
370551	*S. pyogenes* serovar M12, strain MGAS9429	1868	1,836,467
370552	*S. pyogenes* serovar M2, strain MGAS10270	1964	1,928,252
370554	*S. pyogenes* serovar M4, strain MGAS10750	1964	1,937,111
160491	*S. pyogenes* serovar M5, strain Manfredo	1736	1,841,271
293653	*S. pyogenes* serovar M1, strain ATCC BAA-947	1840	1,838,554
160490	*S. pyogenes* serovar M1, strain ATCC 700294	1691	1,852,441
186103	*S. pyogenes* serovar M18, strain MGAS8232	1835	1,895,017
319701	*S. pyogenes* serovar M28, strain MGAS6180	1884	1,897,573
198466	*S. pyogenes* serovar M3, strain ATCC BAA-595	1858	1,900,521
193567	*S. pyogenes* serovar M3, strain SSI-1	1852	1,894,275
286636	*S. pyogenes* serovar M6, strain ATCC BAA-946)	1879	1,899,877
471876	*S. pyogenes* NZ131	1700	1,815,785
388919	*S. sanguinis* SK36	2269	2,388,435
391295	*S. suis* 05ZYH33	2179	2,096,309
391296	*S. suis* 98HAH33	2179	2,095,698
264199	*S. thermophilus* ATCC BAA-250 / LMG 18311	1577	1,796,846
322159	*S. thermophilus* ATCC BAA-491 / LMD-9	1704	1,856,368
299768	*S. thermophilus* CNRZ 1066	1590	1,796,226

Table A.14: Complete list of *Vibrio* isolates used.

TaxID	Isolate	Proteins (#)	Genome size (bp)
243277	*V. cholerae* serovar O1, strain ATCC 39315	3784	4,033,464
345073	*V. cholerae* serovar O1, strain ATCC 39541	3772	4,132,319
312309	*V. fischeri* ATCC 700601 / ES114	3814	4,227,869
388396	*V. fischeri* MJ11	4034	4,323,877
338187	*V. harveyi* ATCC BAA-1116 / BB120	5608	5,969,369
223926	*V. parahaemolyticus* RIMD 2210633	4821	5,165,770
216895	*V. vulnificus* CMCP6	4473	5,126,797
196600	*V. vulnificus* YJ016	4990	5,211,578

Table A.15: Complete list of *Xanthomonas* isolates used.

TaxID	Isolate	Proteins (#)	Genome size (bp)
190486	*X. axonopodis* pathovar citri, strain 306	4354	5,175,554
314565	*X. campestris* pathovar campestris, strain 8004	4239	5,148,708
340	*X. campestris* pathovar campestris, strain B100	4410	5,079,002
456327	*X. campestris* pathovar vesicatoria, strain 85-10	4628	5,178,466
190485	*X. campestris* campestris, strain ATCC 33913	4127	5,076,188
342109	*X. oryzae* pathovar oryzae, strain MAFF 311018	4204	4,940,217
360094	*X. oryzae* pathovar oryzae, strain PXO99A	4587	5,240,075
291331	*X. oryzae* oryzae, strain KXO85 / KACC10331	4380	4,941,439

Table A.16: Complete list of *Yersinia* isolates used.

TaxID	Isolate	Proteins (#)	Genome size (bp)
393305	*Y. enterocolitica* serovar O:8, strain 8081	4021	4,615,899
229193	*Y. pestis* biovar Mediaevalis, strain 91001	4013	4,595,065
187410	*Y. pestis* biovar Mediaevalis, strain KIM5	3968	4,600,755
214092	*Y. pestis* biovar Orientalis, strain CO-92	3908	4,653,728
386656	*Y. pestis* Pestoides F	3942	4,517,345
360102	*Y. pestis* bv., strain Antiqua	4135	4,702,289
349746	*Y. pestis* bv. Antiqua, strain Angola	3821	4,504,254
377628	*Y. pestis* bv. Antiqua, strain Nepal516	3946	4,534,590
273123	*Y. pseudotuberculosis* serovar I, strain IP32953	4016	4,744,671
502801	*Y. pseudotuberculosis* serovar IB, strain PB1/+	4213	4,695,619
349747	*Y. pseudotuberculosis* serovar O:1b, strain IP 31758	4305	4,723,306
502800	*Y. pseudotuberculosis* serovar O:3, strain YPIII	4171	4,689,441