

ABSTRACT

Title of Dissertation / Thesis: COMPUTATIONAL ANALYSES OF MICROBIAL GENOMES – OPERONS, PROTEIN FAMILIES AND LATERAL GENE TRANSFER.

Yongpan Yan, Doctor of Philosophy, 2005

Dissertation / Thesis Directed By: Professor John Moulton, Center for Advanced Biotechnology Research, University of Maryland Biotechnology Institute

As a result of recent successes in genome scale studies, especially genome sequencing, large amounts of new biological data are now available. This naturally challenges the computational world to develop more powerful and precise analysis tools. In this work, three computational studies have been conducted, utilizing complete microbial genome sequences: the detection of operons, the composition of protein families, and the detection of the lateral gene transfer events.

In the first study, two computational methods, termed the Gene Neighbor Method (GNM) and the Gene Gap Method (GGM), were developed for the detection of operons in microbial genomes. GNM utilizes the relatively high conservation of order of genes in operons, compared with genes in general. GGM makes use of the relatively short gap between genes in operons compared with that otherwise found between adjacent genes. The two methods were benchmarked using biological

pathway data and documented operon data. Operons were predicted for 42 microbial genomes. The predictions are used to infer possible functions for some hypothetical genes in prokaryotic genomes and have proven a useful adjunct to structure information in deriving protein function in our structural genomics project.

In the second study, we have developed an automated clustering procedure to classify protein sequences in a set of microbial genomes into protein families. Benchmarking shows the clustering method is sensitive at detecting remote family members, and has a low level of false positives. The aim of constructing this comprehensive protein family set is to address several questions key to structural genomics. First, our study indicates that approximately 20% of known families with three or more members currently have a representative structure. Second, the number of apparent protein families will be considerably larger than previously thought: We estimate that, by the criteria of this work, there will be about 250,000 protein families when 1000 microbial genomes are sequenced. However, the vast majority of these families will be small. Third, it will be possible to obtain structural templates for 70 – 80% of protein domains with an achievable number of representative structures, by systematically sampling the larger families.

The third study is the detection of lateral gene transfer event in microbial genomes. Two new high throughput methods have been developed, and applied to a set of 66 fully sequenced genomes. Both make use of a protein family framework. In the High Apparent Gene Loss (HAGL) method, the number and nature of gene loss events

implied by classical evolutionary descent is analyzed. The higher the number of apparent losses, and the smaller the evolutionary distance over which they must have occurred, the more likely that one or more genes have been transferred into the family. The Evolutionary Rate Anomaly (ERA) method associates transfer events with proteins that appear to have an anomalously low rate of sequence change compared with the rest of that protein family. The methods are complementary in that the HAGL method works best with small families and the ERA method best with larger ones. The methods have been parameterized against each other, such that they have high specificity (less than 10% false positives) and can detect about half of the test events. Application to the full set of genomes shows widely varying amounts of lateral gene transfer.

COMPUTATIONAL ANALYSES OF MICROBIAL GENOMES – OPERONS,
PROTEIN FAMILIES AND LATERAL GENE TRANSFER

By

Yongpan Yan

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2005

Advisory Committee:
Professor John Moulton, Chair
Professor Raymond J. St. Leger
Associate Professor Stephen M. Mount
Assistant Professor Jocelyne DiRuggiero
Associate Professor Chau-Wen Tseng

© Copyright by
Yongpan Yan
2005

DEDICATION

To my parents

ACKNOWLEDGEMENTS

My foremost thanks go to my research advisor, Dr. John Moulton. I still remember six years ago when I started my rotation in John's group, I knew nothing about computational biology. He led me into this fascinating new world and gave me endless inspiration and encouragement, especially when I encountered research difficulties. He is truly an excellent mentor. Without him, I can't imagine how I achieved so much after six years' study

I would also like to thank Mr. Eugene Melamud. Being a Bioinformatics and computer 'geek' (according to himself), he gave me a lot of help. I am also grateful to my colleagues, Mr. Peng Yue, Dr. John Collins, Mr. Zhen Shi, Mr. Ethan Zhang, Dr. Carol Scott, Dr. Zheng Wang, Mr. Holger Klein, Dr. Chia-en Chang and many others in CARB. They all showed their kind support during my study.

I also would like to express my appreciation to my committee members: Dr. Steve Mount, Dr. Jocelyne Diruggiero, Dr. Chau-wen Tseng and Dr. Raymond St leger. They are knowledgeable, professional and kind. They have given me a lot of excellent advice. This work won't be the same without their help and encouragement.

I also own great thank to my family, Mom, Dad and my brother. Although far from here, they have given me the priceless support. I know their hearts are always with me. My special thank is for my wife. Thank her for always being there to support me.

Table of Contents

DEDICATION	ii
ACKNOWLEDGEMENTS	iii
Table of Contents	iv
List of Tables	vi
List of Figures	vii
Chapter 1: Introduction	1
1.1 Motivation	1
1.2 Background	5
1.2.1 Operon	5
1.2.2 Protein Families for Structural Genomics	5
1.2.3 Lateral Gene Transfer	6
1.3 Overview	7
Chapter 2: Detection of Operons	8
2.1 Introduction	8
2.2 Methods	12
2.3 Results	25
2.3.1 Gene Neighbor Method	25
2.3.2 Gene Gap Method	30
2.3.3 Methods Comparison	34
2.3.4 Prediction of Protein Function	37
2.3.5 Combination of Structure and Operon Information	44
2.3.6 Extent of Operon Conservation and implications for Pathway Consistency	48
2.4 Conclusion and Discussion	50
Chapter 3: Protein Family Clustering for Structural Genomics	55
3.1 Introduction	55
3.2 Methods	58
3.3 Results	71
3.3.1 Protein Family Clustering	71
3.3.2 Structural Genomics Analysis	79
3.4 Conclusion and Discussion	93
Chapter 4: Lateral Gene Transfer between Prokaryotic Genomes	100
4.1 Introduction	100
4.2 Methods	106

Generation of a Domain-based Protein Family Set	110
Phylogenetic Tree Construction.....	112
4.3 Results.....	127
4.3.1 Phylogenetic Tree	127
4.3.2 The High Apparent Gene Loss method (HAGL).....	131
4.3.3 The Evolutionary Rate Anomaly method (ERA).....	134
4.3.4 Calibration and Evaluation of the Methods	144
4.3.5 Application to the Set of 66 Genomes	148
4.4 Conclusion and Discussion	153
Chapter 5 Conclusions	158
Bibliography	160

List of Tables

2.1 Gene pair conservation for the <i>histidine</i> operon	26
2.2 An example of a predicted operon in <i>H.influenzae</i>	39
2.3 The number of genes whose functions can be predicted	43
2.4 Function predictions using operon information in S2F.....	47
3.1 The number of proteins in 67 fully sequenced microbial genomes.....	60
3.2 The number of proteins in 73 recently sequenced genomes.....	62
3.3 Agreement between generated and PfamA family, as a function of PSI-BLAST E score threshold	74
3.4 Agreement between generated and PfamA family, as a function of clustering procedures	76
4.1 66 fully sequenced microbial genomes used in the LGT analysis	109
4.2 14 well conserved orthologous protein families	113
4.3A Calibration and Evaluation of the ERA method	146
4.3B Calibration and Evaluation of the HAGL mehod	148
4.4 LGT predictions in microbial genomes.....	153

List of Figures

2.1	Conserved gene pair	13
2.2A	Common Neighbor Fraction (CNF) phylogenetic tree	16
2.2B	Common Gene Fraction (CGF) phylogenetic tree	18
2.2C	16S ribosomal RNA phylogenetic tree.....	20
2.3	Gene order in the <i>E.coli</i> <i>K12</i> histidine operon	26
2.4A	Specificity of the GNM as a function of the conservation level.....	28
2.4B	Sensitivity of the GNM as a function of the conservation level	29
2.5	Example of an <i>E.coli</i> operon conforming to the expectations of the GGM...	30
2.6A	Distribution of intergene gap lengths in <i>E.coli</i>	32
2.6B	Specificity of the GGM in <i>E.coli</i> as a function of the maximum gap length allowed.	33
2.6C	Sensitivity of the GGM in <i>E.coli</i> as a function of the maximum gap length	34
2.7	Comparison of operon gene pair predictions by the two methods	36
2.8	Histogram of the conservation level of predicted operon gene pairs in <i>E.coli</i>	49
2.9	Number of predicted <i>E.coli</i> operons as a function of operon size (number of genes) and conservation level of the complete operon	50
3.1	An example of domain parsing	63
3.2	Domain merging check	66
3.3A	Distribution of domain family size.....	72
3.3B	Number of Singletons, Doubletons and all others and the percentage of sequence space covered by each of the three categories.	73
3.4	Benchmarking of the family building procedure.....	78

3.5	Family structure coverage as a function of family size.....	80
3.6A	Number of families as a function of the number of genomes	83
3.6B	Log-log view of the relationship between the number of families and number of genomes considered.....	84
3.6C	Predicted number of families as a function of the number of fully sequenced prokaryotic genomes.....	85
3.6D	Predicted number of apparent singletons as a function of the number of fully sequenced prokaryotic genomes	86
3.7	Cumulative number of experimental structures needed to obtain complete coverage of families size 3 and larger	87
3.8A.	Number of families with representative structures needed to provide structural coverage for different fractions of protein domains, as a function of the number of fully sequenced genomes.....	90
3.8B.	Projection of the number of families with representative structures needed to obtain structural coverage of different fractions of protein domains ...	91
3.8C.	Expansion of Figure 3.8B for coverage between 50% and 80%.....	92
4.1	Phyletic pattern of a hypothetical protein family in a species tree.....	115
4.2	Example of Evolutionary Rate analysis of an Orthologous Protein Family	119
4.3A	The Neighbor Joining tree for 66 microbial genomes, derived from 14 conserved protein families	129
4.3B	16S ribosomal RNA Neighbor Joining tree for the 66 genomes	130
4.4:	Distribution of protein families as a function of the number of apparent gene loss events in each family and family size	132
4.5.	Distribution of T (ratio of losses to total branch length) for the 4856 orthologous protein families with three or more members	134
4.6A.	Distribution of evolutionary rates for 4116 orthologous protein families	136
4.6B.	Distribution of $\sigma[(R(i,j)-\langle R(i,j)\rangle)/\langle R(i,j)\rangle]$, the relative standard deviation of evolutionary rates within protein families	137

4.7A. Relative rates of amino acid substitution between pairs of proteins in the mercuric resistance operon regulatory protein family (MerR).....	139
4.7B. The phylogenetic and species trees of the mercuric resistance operon repressor protein (MerR) family	140
4.8A. Example of a protein family with anomalously high rates of amino acid change for one of its members	141
4.8B. The phylogenetic tree of a putative endonuclease protein family	143

Chapter 1: Introduction

1.1 Motivation

The explosion of knowledge of genome sequences offers us tremendous new opportunities for addressing questions of major biological interest. However, most of the corresponding proteins have not been experimentally characterized. The challenge of understanding the role of these proteins has led to the development of a range of functional genomics methods, which generate various types of information ¹.

Structural Genomics is one of such approach which aims to provide structure for a high fraction of natural proteins. The general intent of Structure Genomics is not to obtain an experimental structure for each protein. Rather, protein sequences are clustered into families, and one or more representative structures are determined for each protein family. Computational comparative modeling is then used to provide model structures for other family members. In this sense, current Structural Genomics is a combined experimental and computational effort.

Protein structure provides a very powerful means of understanding aspects of protein function. Generally, during evolution, structure is well conserved and can be used to detect remote evolutionary relationships between proteins that are often not detectable by current sequence alignment methods. Therefore, when the structure of a

'hypothetical' protein is obtained experimentally, its newly revealed structural homologues can sometimes be used to help identify the function. In cases where a function cannot be identified by homology, clues such as potential active sites and the location of conserved residues can provide a starting point for more conventional methods of function determination.

In our center (Center for Advanced Research in Biotechnology, CARB), a structural genomics project (<http://s2f.umbi.umd.edu/>) was initiated in 1998. The initial goal was to determine structures of hypothetical proteins from *Haemophilus influenzae* and then combine structure information with computational analysis to predict the protein function. So in my first project – detection of operons in microbial genomes, I have developed two methodologies to identify operons, with the aim of providing function clues for hypothetical proteins, using the strong relationship between the operon structure and function relatedness. Prediction results were successfully utilized to infer the possible biological function of some hypothetical proteins.

The second project - protein family clustering for structural genomics, aims to provide a complete and reliable set of families for all the proteins in a set of microbial organisms. A multi-linkage clustering scheme was developed to facilitate family construction. The families were thoroughly benchmarked using SCOP^{2;3} and PFAM⁴ data, and it was shown that the clustering method is more sensitive in detecting remote evolutionary relationships than other alignment methods, when the false positive rate is low. The completeness of this set makes it possible to obtain

improved estimates of the number and diversity of families in the prokaryotic kingdom. Several important questions related to structural genomics strategy were addressed using this family set: (1) What is the structure coverage for currently known families? (2) How will the number of known apparent families grow as more genomes are sequenced? (3) What is a practical strategy for maximizing structure coverage in future? Our study indicates that approximately 20% of known families with three or more members currently have a representative structure. The number of apparent protein families will be considerably larger than previously thought: The estimate is that, by the criteria of this work, there will be about 250,000 protein families when 1000 microbial genomes have been sequenced. However, the vast majority of these families will be small, and it will be possible to obtain structural templates for 70 – 80% of protein domains with an achievable number of representative structures, by systematically sampling the larger families.

The third project, detection of lateral gene transfer in microbial genomes, originated with the idea of estimating the age distribution of protein folds, a fundamental question in Structural Genomics. Sequence analysis of bacterial genomes reveals a large number of apparent singletons – proteins found in one organism, but with no detectable relatives in any other organism^{5; 6; 7}. These proteins, together with a large number of protein families that appear to have members in a very few genomes, dominate protein family space, and suggest that new protein folds may arise frequently in evolution. To test this possibility, we began investigating the apparent age of protein families. It quickly became clear that no analysis of family age is

possible without taking into account the occurrence of lateral gene transfer. Further, LGT is in itself a central issue in the evolution of bacteria.

Thus, the LGT project became a project on its own right. In this study, we have developed two methods to observe lateral gene transfer events in genome scale. The High Apparent Gene Loss (HAGL) method detects LGT events by counting the minimum number of losses necessary to explain the phyletic pattern of a protein family. LGT events are likely when there are a large number of losses in a family over a small evolutionary distance. The Evolutionary Rate Anomaly (ERA) method finds LGT events by identifying genes which have statistically different rates of sequence change from the family average. Because of the different signals utilized, each method detects a largely different set of LGT events. The two methods combined do not cover the whole spectrum of LGT possibilities, but do provide a useful sampling, and confirm that LGT is very widespread. Grouping laterally transferred genes in terms of genomes and families shows the distribution terms are uneven.

1.2 Background

1.2.1 Operon

An operon is a set of adjacent genes that share the same regulatory machinery and are transcribed into a single mRNA molecule. A well known example is the lac operon in *Escherichia coli*. This operon includes three genes: *lacA*, *lacY* and *lacZ*, sharing the same promoter and terminator, so that transcription produces a single mRNA molecule. Operons are widespread in prokaryotic organisms and are also found in some eukaryotes such as *C.elegans*. An important characteristic is that genes in an operon are very likely to have related functions. For example, the three genes in the lac operon all play roles in the lactose metabolic pathway in *E.coli*. For this reason, we can use a predicted operon to infer the function of hypothetical proteins, when other genes in the same operon have clear function annotation.

1.2.2 Protein Families for Structural Genomics

As we explained in the Motivation, the ultimate goal of structural genomics is to provide structures for all biological proteins. Although there have been enormous improvements in experimental methods for determining structure, these still lag behind the genome sequencing by orders of magnitude. As a result, currently only about 1% of proteins with known sequence also have an experimentally known structure. Structural genomics proposes to efficiently provide structural coverage for

proteins by experimentally determining one representative structure for each family and using comparative modeling methods to obtain model structures for other family members. To implement this strategy, a set of protein families is required.

Many protein classification schemes have already been developed, for various purposes. For example, SCOP^{2;3} and CATH⁸ classify proteins in terms of the structural similarity. Pfam⁴ groups proteins based on Hidden Markov Model sequence profiles. None of these classifications is ideal for structural genomics, and an automated procedure which can classify all of known sequence space is needed.

1.2.3 Lateral Gene Transfer

Lateral gene transfer, also called horizontal gene transfer, is the process of transfer of genetic information between different species. The significance of lateral gene transfer was not appreciated until the 1950s, when resistance to penicillin class antibiotics spread rapidly through many pathogens as a result of plasmid transfer⁹. As we now know, two processes act together to shape genome in the evolution of prokaryotic organisms: direct gene inheritance and lateral gene transfer. However, for a long time, it was commonly believed that lateral gene transfer was rare and the dominant process in evolution is inheritance. With the development of genome sequencing, it becomes more and more obvious that lateral gene transfer is also a very significant force.

1.3 Overview

The dissertation is organized as follows. Chapter 2 describes the detection of operons in microbial genomes. Two prediction methods were developed and the results were benchmarked using function pathway data and documented operon data. The prediction results were used to provide function clues for some of the hypothetical proteins in these organisms. Chapter 3 describes the generation of a set of protein families which classify all the proteins in a set of microbial genomes. The family set was thoroughly benchmarked with SCOP data and PFAM data. This set of families was used to answer several important questions of structural genomics. Chapter 4 describes the detection of lateral gene transfer in microbial genomes. Two methods were developed, the HAGL method (High Apparent Gene Loss) and the ERA method (Evolutionary Rate Anomaly). In the absence of an experimental gold standard to use as a benchmark, the prediction results of the two methods were compared. The results were grouped in terms of genomes. Chapter 5 summarizes the conclusions of the three projects and discusses prospects for further work in these areas.

Chapter 2: Detection of Operons

2.1 Introduction

Knowledge of the complete genome sequences of multiple prokaryotic organisms provides many opportunities for analysis.¹⁰ Here we focus on the detection of operons, sets of adjacent genes that share the same regulatory machinery and are transcribed into a single mRNA molecule. Operons are widespread and frequent in prokaryotic organisms, and are also found in some eukaryotes such as *C. elegans*^{11;}
¹².

Analysis of operons of known function, primarily in *E.coli*, has established that genes in an operon are very likely to have related functions^{13; 14; 15}. For this reason, a number of operon detection methods have already been developed. One approach uses nucleotide sequence patterns that are conserved across multiple genomes to identify gene expression regulatory sites, such as promoters and terminators, and so find transcriptional units^{16,17}. These sequence motifs are short and can be highly variable, limiting the prediction capability of this method. The method can be extended by also considering conservation of functional class within operons¹⁸.

Machine learning methods have been used to develop predictive models based on a variety of information, including sequence data, gene expression data, and functional annotations associated with genes ¹⁹. These authors built separate models for the prediction of promoters, terminators and operons. The separate predictions were then combined with a dynamic programming method to map every known and putative gene in a given genome into its most probable operon. The full power of this method is only applicable to very well-studied systems such as *E. coli*.

An alternative approach, which we have also used, is to search for pairs of homologous genes that are adjacent in multiple, phylogenetically well separated prokaryotic genomes ^{20; 21; 22}. As discussed later, gene shuffling is relatively rapid during evolution, so that a strong selective force is required to maintain gene order over long periods of evolutionary time. Operons provide such a selective pressure ^{23; 24}. Thus, conservation of gene pairing across many genomes is evidence of operon structure. Several different algorithms have been developed to make use of this signal. Overbeek et al. ¹⁴ considered all gene pairs conserved in at least two of 30 included genomes to be in operons. The results were used to predict the function of some genes by association with their conserved neighbors. This pioneering method was not benchmarked in any way, however. Wolf and colleagues ²⁵ developed a local gene-by gene genome alignment method similar to a sequence local alignment tool. Pairs of genes aligned across two genomes are assigned a score of 1, and all other gene pairs are assigned a score of zero. All consecutive runs of two or more scores of '1' are then considered to form operons. The statistical significance of the local

alignments was evaluated by using Monte Carlo simulations to obtain an estimate of the random expectation for each score value. The method provided the first measure of prediction reliability. Ermolaeva et al.²¹ developed a method which estimates the likelihood that a conserved gene pair could be in an operon and benchmarked the result with the RegulonDB data set. The method uses an ingenious but indirect statistical model, and has relatively low sensitivity, in the 30 – 50% range.

Yet another approach is to detect operons on the basis of a short inter-gene gap between neighboring genes. Salgado et al.²⁶ combined this approach with function class information to predict operons in *E.coli*. We have also developed a version of this method.

The work reported here builds on the earlier results for the gene neighbor and gene gap models, using a larger set of genomes, combining the conserved neighbor and inter-gene gap methods, and carefully benchmarking the results against two sources. One benchmark makes use of the fact that genes in the same operon are very likely to have related functions, and therefore to belong to the same KEGG²⁷ pathway, allowing us to determine the specificity of the methods. That is, the fraction of genes predicted to be in operons that are correct. The second benchmark allows us to determine the sensitivity of the method, that is, the fraction of operons in a genome that are detected, using known *E.coli* operons in the RegulonDB database²⁸.

With optimum parameters, the specificity of the gene neighbor method is 93% and the sensitivity is 70%. For the gene gap method the specificity is 95% and the sensitivity is 68%. 87% of all operons in the test set are predicted by one or both methods. A higher specificity of 98% is obtained by considering only those operon predictions for which the methods agree, at the expense of a lower sensitivity of 50%.

We have used both methods to predict operons in a set of 42 microbial genomes, and thus provide some indications of function for a large number of hypothetical proteins. The primary motivation for the work and major use of the results is to obtain insight into the function of previously unannotated proteins in *Haemophilus influenzae* and *E.coli*, by combining operon prediction with information from structure obtained as part of a structural genomics project (<http://s2f.umbi.umde.edu>).

2.2 Methods

Database: All downloaded and generated information was stored in a MySQL relational database running on a Linux server.

Genome Data: The complete genome sequences together with open reading frame annotations were retrieved from the NCBI genome sequence database (<http://www.ncbi.nlm.nih.gov/Genomes/index.html>).

Homolog detection: The sequence of each protein in the set of genomes was compared with all others, using three rounds of PSI-BLAST^{29; 30}. A pair of proteins A and A' in two different organisms are considered homologous if there is a PSI-BLAST hit with an E-value of 0.001 or lower for A' when A is used as the search sequence, and for A, using A' as the search sequence.

Adjacent Gene Pair: Any pair of adjacent genes on the same strand in a genome is defined as an adjacent gene pair.

Conserved Gene Pair: an adjacent gene pair is considered to be conserved across two genomes if the following conditions hold:

1. Genes A and B form an adjacent gene pair in Genome 1 and genes A' and B' form an adjacent gene pair in Genome 2.
2. Genes A' and B' are homologs of genes A and B respectively.

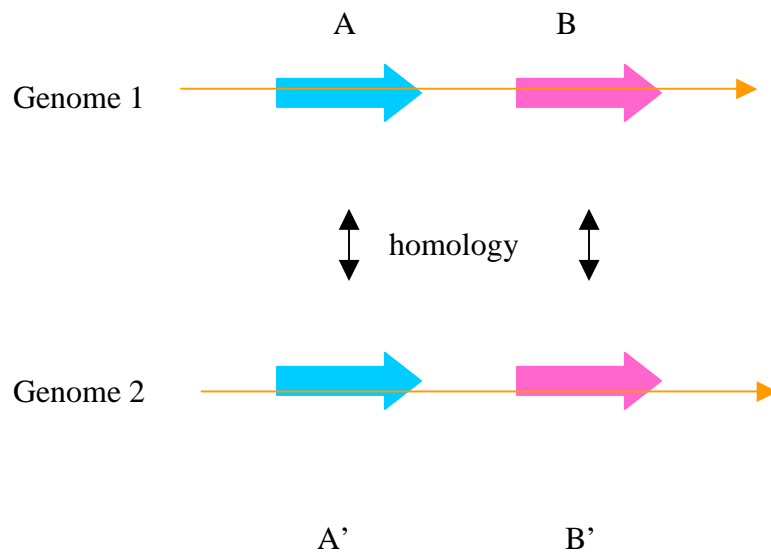


Figure 2.1. Conserved gene pair. A and A' and B and B' are pairs of homologous proteins. A and B form an adjacent gene pair in Genome 1 and A' and B' form an adjacent gene pair in Genome 2.

Conserved gene pairs may be observed for three possible reasons: (1) The genes are part of an operon present in both genomes – the signal we seek to detect. (2) Insufficient time has elapsed since the genomes diverged for gene shuffling to be complete. We greatly reduce these incidences by only comparing well diverged genomes. (3) The pair order occurs by chance. For typical genomes, with more than 1000 genes, the probability of chance pairing across two genomes is less than 0.001.

Genome Separation: The relative divergence of pairs of genomes is measured in terms of the completeness of gene shuffling, using the Common Neighbor Fraction (CNF) within a genome pair.

$$\text{CNF} = \text{Number of conserved gene pairs} / \text{genome size}$$

where ‘genome size’ is the number of genes in the smaller of the two genomes compared. Two well shuffled genomes will have a low CNF value, but greater than zero primarily because of the effect of neighbor conservation in operons, while recently diverged or slowly shuffling ones will have a larger value.

Common Neighbor Fraction Tree: An inter-genome distance matrix was constructed for the 42 genomes, with elements

$$D(I,J) = [1 - \text{CNF}(I,J)]$$

where $\text{CNF}(I,J)$ is the common neighbor fraction between genomes I and J. A neighbor joining tree was built from this distance matrix, using the NEIGHBOR program in PHYLIP package (Felsenstein 1989)³¹, and using *Saccharomyces cerevisiae* as the outgroup. The resulting tree is shown in Figure 2.2A. The more shuffled a pair of genomes with respect to each other, the further up the tree the branch point between them. Shuffling was considered incomplete for genomes linked by branch points above the vertical green line. Each set of genomes diverging from

one of the 30 branches crossing the green line was defined as a single **Genome Group**.

Common Gene Fraction Tree: A Common Gene Fraction (CGF) tree was built using an inter-genome distance matrix composed of elements $[1 - \text{CGF}(I,J)]$, where $\text{CGF}(I,J)$ is the common gene fraction between genomes I and J, and:

$$\text{CGF} = \text{Number of homologous pairs} / \text{genome size}$$

This tree, shown in figure 2.2B, represents the relationship between genomes in terms of the number of detectable homologous gene pairs between them.

16S ribosomal RNA Tree: A standard 16S ribosomal RNA neighbor joining tree³² was also constructed, and is shown in Figure 2.2C. 16S ribosomal RNA data was retrieved from NCBI Genbank. The RNA distance matrix was built using the DNADIST program in the PHYLIP package (Felsenstein 1989)³¹. This tree represents the relationship between genomes in terms of sequence conservation. Properties of all the trees are discussed in the Results section.

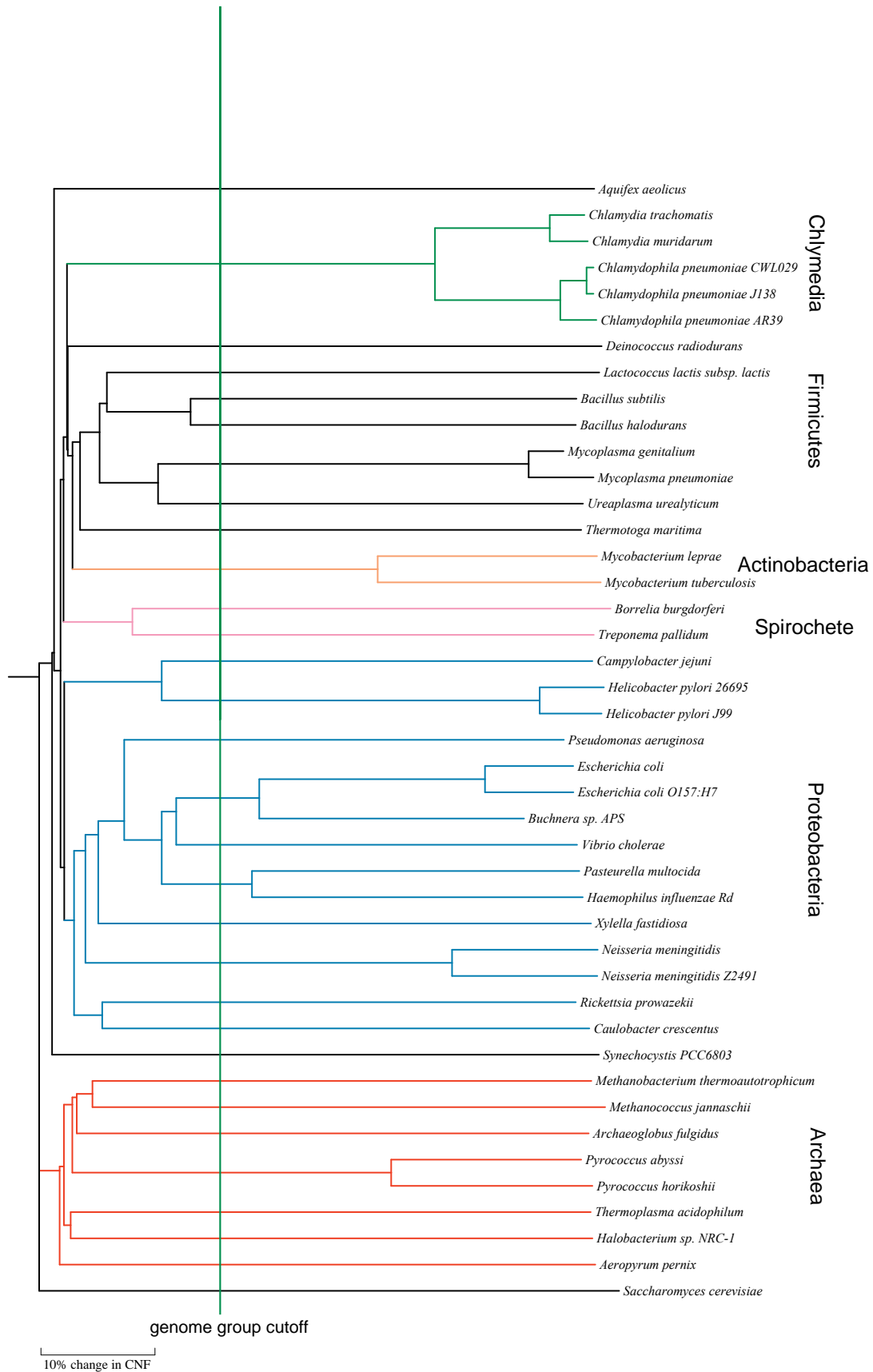


Figure 2.2A. Common Neighbor Fraction (CNF) phylogenetic tree for 42 fully sequenced prokaryotic genomes. Relationships in the tree are determined by the

matrix of $[1-\text{CNF}(I,J)]$ values, where $\text{CNF}(I,J)$ is the common neighbor fraction between genomes I and J – a measure of the extent of gene order conservation. On this basis, genomes are segregated into the standard large groups such as Archaea, Proteobacteria, Firmicutes and so on. Each of these groups is shown in a different color. The relative closeness of most branch-points to the tree root is an indication of rapid shuffling of gene order during evolution of these organisms. For operon identification purposes, pairs of genomes with branch points above the vertical green line were considered incompletely shuffled. The tree was built using the neighbor joining method, with *Saccharomyces cerevisiae* as the outgroup.

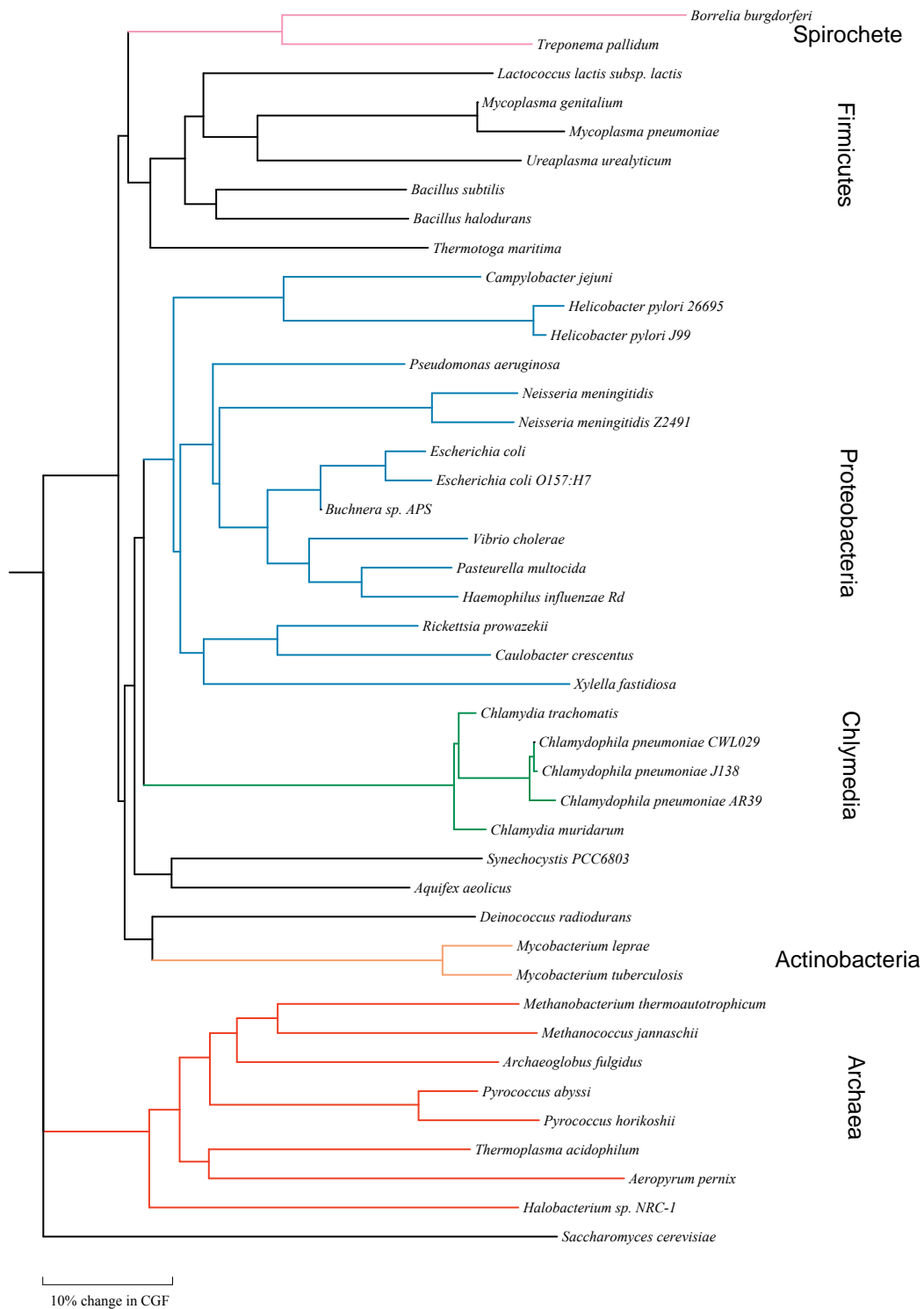


Figure 2.2B. Common Gene Fraction (CGF) phylogenetic tree for the same set of genomes as figure 2.2A. Here, relationships in the tree are determined by the value $[1 - \text{CGF}(I, J)]$ where $\text{CGF}(I, J)$ is the common gene fraction between genomes I and J –

a measure of the extent of gene homology. The same major subgroupings as in figure 2.2A are observed, but branch points are usually further from the root, indicating that apparent gene gain and loss are slower processes than gene shuffling. The tree was built in the same manner as that in figure 2.2A, with the same color scheme, indicating the major microbial genome groups.

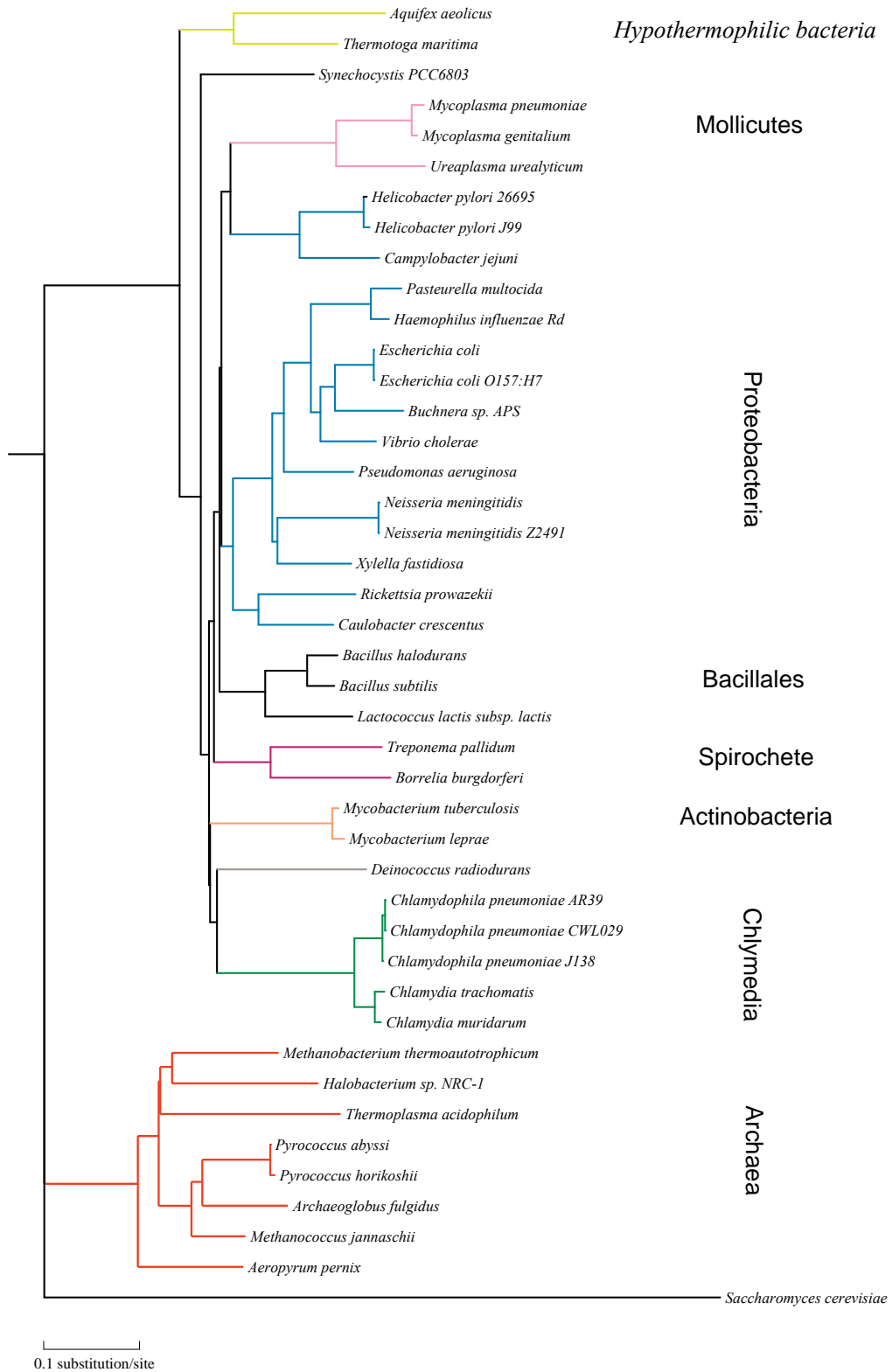


Figure 2.2C. 16S ribosomal RNA phylogenetic tree for the same set of genomes as figures 2.2A and 2.2B. Here, relationships in the tree are determined by the extent of sequence identity in 16S ribosomal RNA between pairs of genomes. The bar shows the scale in units of accepted base substitutions. The same major subgroupings as in 2a and 2b are observed, but the positions of branch points are farther away from the root than in the other trees. The tree was built in the same manner as that in figure 2.2A, with the same color scheme, indicating major microbial genome groups.

Conservation level: As noted above, the set of 42 genomes is divided into 30 genome groups. The conservation level of a conserved gene pair is defined as the number of genome groups in which one or more instances of the pair are found. The higher the conservation level, the more likely that the gene pair is in an operon.

Inter-gene Gap: The length of the non-coding region between two genes which are adjacent to each other on the same strand is obtained by subtracting the position of the last base in the first gene from the position of the first base in the second gene, using NCBI nucleotide indexing. Since two adjacent genes may have overlapping coding regions, this length can be negative.

Operon gene pair: Any adjacent gene pair considered to be part of operon is termed an operon gene pair. Presence in an operon is defined by a conservation level above a specified threshold or an inter-gene non-coding region length below some threshold, or by both.

Accuracy measures: The accuracy of the two operon prediction methods is expressed in terms of specificity (fraction of true negatives correctly identified in a test set) and sensitivity (fraction of true positives correctly identified in a test set). I.e.

$$\text{Specificity (Sp)} = \text{TN} / (\text{TN} + \text{FP})$$

Where TN is the number of true negatives in a test set, FP is the number of false positives, and (TN + FP) is the total number of points in the set.

$$\text{Sensitivity (Sn)} = \text{TP} / (\text{TP} + \text{FN})$$

Where TP is the number of true positives in a test set, FN is the number of false positives, and (TP + FN) is the total number of points in the set.

Test sets: Constructing suitable test sets is a key component of the statistical evaluation of any prediction method. Ideally, one test set should be used for calculating all accuracy measures. No such comprehensive test set exists for operon evaluation, so we have used separate sets for measuring specificity and sensitivity. For specificity, we require a test set where true negatives and false positives can be counted. That is, a set where we have knowledge of which pairs of genes are not in the same operon, but not necessarily knowledge of which pairs of genes are in the same operon. We have used data from KEGG pathways for this. KEGG (Kyoto

Encyclopedia of Genes and Genomes) (<http://www.genome.ad.jp/kegg/kegg2.html>)^{33; 34; 35} contains hand-curated data from the literature, identifying which genes are in the same functional pathway. The data are most extensive and reliable for *E.coli K-12*, and the genes in this organism that are assigned to a KEGG pathway form the basis of the test set. We assume that any adjacent gene pair where the genes are assigned to different KEGG pathways cannot be in the same operon. Therefore, any adjacent gene pair predicted to be in an operon and for which the member genes are assigned to different KEGG pathways is counted as a false positive. While this definition is imperfect (see below), it is likely to result in an over-estimate of false positives, rather than an under-estimate. Conversely, any adjacent gene pair for which the genes are assigned to different KEGG pathways and which is not predicted to be in an operon is counted as a true negative. Thus, true negatives (TN) and the false positives (FP) are given by:

TN = number of adjacent gene pairs assigned to different KEGG pathways and not predicted to be in the same operon.

FP = number of adjacent gene pairs assigned to different KEGG pathways and predicted to be in the same operon.

There are a total of 4287 adjacent gene pairs in the *E.coli K-12* genome¹⁸, of which 599 have both member genes assigned to KEGG pathways. 161 of these pairs are assigned to different pathways, and form the potential test set. A check of these gene

pairs against the literature revealed that, contrary to expectation, 35 of the 161 are known to be in the same operon^{28; 36; 37; 38; 39; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 50; 51; 52; 53; 54; 55;}

⁵⁶ These gene pairs were removed, leaving 126 pairs for the specificity test. We also observed 438 cases where adjacent gene pairs are in the same KEGG pathway. However, two genes may be in the same pathway without being in the same operon, so these data are not suitable for use in a specificity assessment.

For sensitivity evaluation, we require a test set where true positives and false negatives can be counted. That is, a test set where we have knowledge of which pairs of adjacent genes are in the same operon, but not necessarily knowledge of which pairs are not in the same operon. RegulonDB²⁸, a collection of experimentally determined *E. coli K-12* operons from the literature, was used as the source of adjacent gene pairs in the same operon (Version 3.2, with 240 operons containing a total of 593 pairs, and in all including 830 genes). The set of adjacent gene pairs form the test set. Any predicted operon gene pair which matches one of these experimentally determined pairs is counted as a true positive, and any experimentally known operon gene pair not predicted is counted as a false negative. That is:

True positives (TP) = number of gene pairs which are in both the RegulonDB experimentally determined operon set and the prediction set.

False negatives (FN) = number of gene pairs which are in the RegulonDB operon set but not in the prediction set.

2.3 Results

2.3.1 Gene Neighbor Method

Gene Shuffling is a relatively rapid process

The Common Neighbor Fraction (CNF) phylogenetic tree, the Common Gene Fraction (CGF) tree and the 16S ribosomal RNA tree are shown in Figure 2. All three trees show the expected approximate topologies. For example, archaea and bacteria are on separate branches and bacteria are divided into Proteobacteria, Gram-positive, Chlamydia, Spirochete and Hyperthermophilic bacteria. The Proteobacteria are further separated into four subdivisions and Gram positive bacteria are separated into two subdivisions. The key difference between these trees is that the branching points in the CNF tree are closer to the root node. This indicates that the rate of gene shuffling is much faster than the rate of loss of orthologs between genomes (captured by the CGF tree) and the rate of sequence change in conserved bio-molecules, such as 16S ribosomal RNA. Similar observations have been made by Bork and colleagues⁵⁷;⁵⁸. Rapid shuffling facilitates a low false positive rate of operon identification.

An example of pair conservation in an operon

Conserved gene pairs were identified using gene neighbor conservation in the 30 genome groups. At a conservation level of 3 or above, *E. coli* has 1077 conserved gene pairs. Figure 2.3 and table 2.1 show an example, the *E. coli K-12 histidine* operon with its gene members *hisG, D, C, B, H, A, F* and *I*. This well established operon⁵⁹ contains eight genes, and hence seven adjacent gene pairs. Table 2.2 shows the conservation level for each of these pairs.

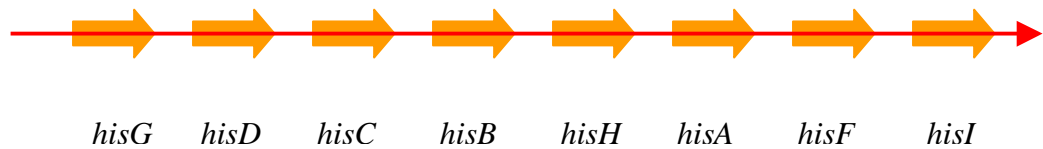


Figure 2.3. Gene order in the *E. coli K-12 histidine* operon.

gene1	Gene2	Conservation level
<i>hisG</i>	<i>hisD</i>	11
<i>hisD</i>	<i>hisC</i>	8
<i>hisC</i>	<i>hisB</i>	7
<i>hisB</i>	<i>hisH</i>	13
<i>hisH</i>	<i>hisA</i>	13
<i>hisA</i>	<i>hisF</i>	20
<i>hisF</i>	<i>hisI</i>	10

Table 2.1. Gene pair conservation for the *histidine* operon. The conservation level for each adjacent gene pair is shown. The genes in this operon belong to the Histidine metabolism pathway in *E.coli* (KEGG pathway eco00340).

All gene pairs within this operon are supported by a substantial conservation level, although the level varies quite widely, from a low of seven to a high of 20 (i.e. adjacent homologs of these pairs are found in between seven and 20 other genome groups). Varied conservation may arise from a number of factors. First, complete pathways may not be conserved across all bacterial genomes. Second, there may be gene reordering within an operon, not tracked by the present method. Third, for rapidly evolving genes, remote orthologs may not be detected.

Specificity and Sensitivity of the Gene Neighbor Method

As described in methods, the specificity is given by:

$$Sp = TN/126$$

Where TN (true negatives) is the number of the 126 *E.coli* adjacent gene pairs in different *E.coli* KEGG pathways that are not predicted to be in the same operon. It is expected that the higher the conservation level required to define an operon gene pair, the higher the specificity. Figure 2.4A shows specificity as a function of the minimum required conservation level.

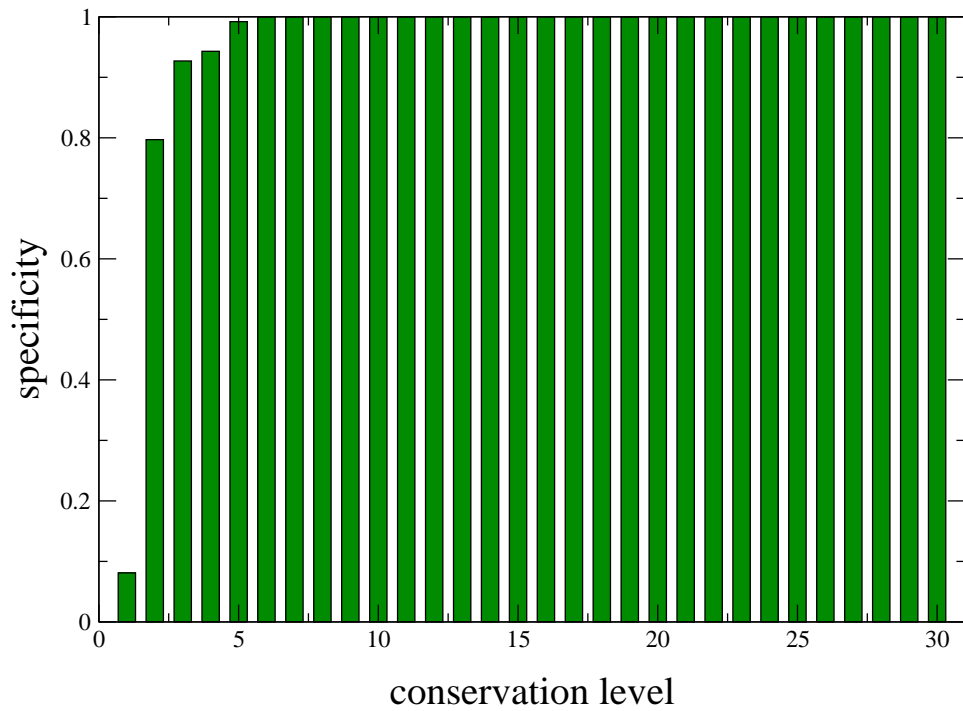


Figure 2.4A. Specificity of the Gene Neighbor Method as a function of the minimum conservation level required. The higher the conservation level, the greater the specificity (i.e the fewer false positives). At a conservation level of three, the specificity is 93%.

Sensitivity is measured as

$$Sn = TP/593$$

Where TP (true positives) is the number of the 593 RegulonDB operon gene pairs identified. We expect the sensitivity to decrease with increase in the conservation level required for an operon gene pair. Figure 2.4B shows the sensitivity as a function of the minimum required conservation level.

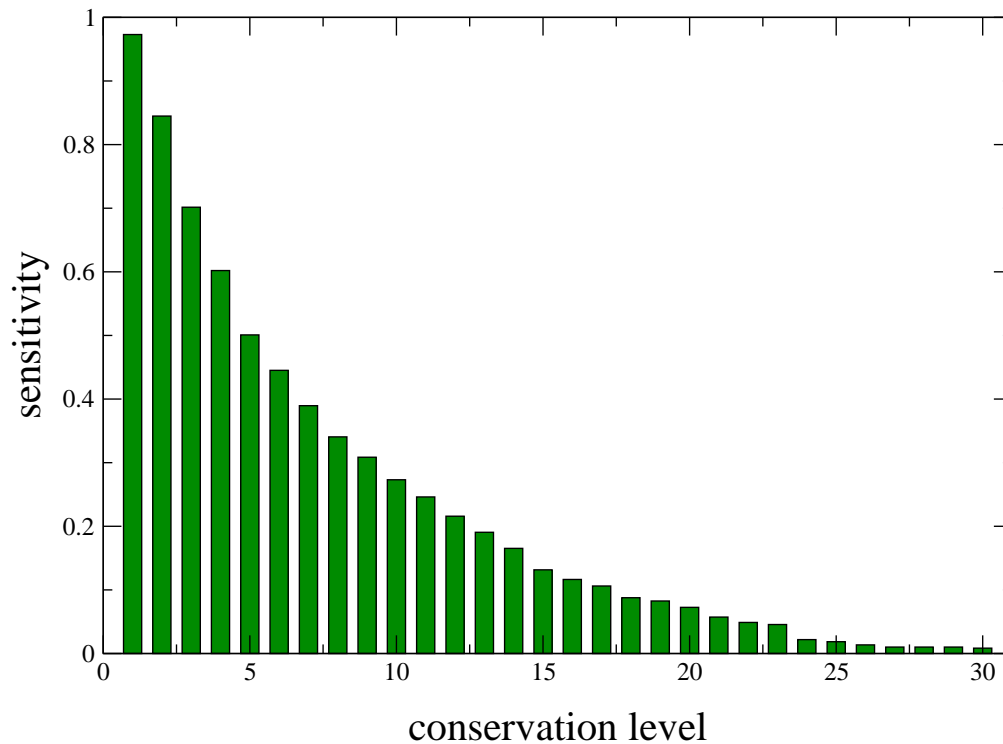


Figure 2.4B. Sensitivity of the Gene Neighbor Method as a function of the minimum conservation level required. As the conservation level increases, the sensitivity falls rapidly (i.e. fewer true operon gene pairs are included). At a conservation level of three, the sensitivity is 70%.

Choice of a higher minimum conservation level will result in fewer incorrect assignments of operon gene pairs, but also fewer of the true pairs will be identified. Based on the data in Figure 2.4, we chose a conservation level of three as the minimum. I.e. an adjacent gene pair must be present in at least three genome groups to be considered part of operon. With this threshold, the specificity is 93% and the sensitivity is 70%. In *E.coli*, with this threshold, 1073 operon gene pairs are predicted. The numbers of predicted operon gene pairs in all 42 microbial genomes are listed in Table 2.3.

2.3.2 Gene Gap Method

The principle of the Gene Gap Method is that a short non-coding region between two genes is too small to hold any regulatory machinery, such as a promoter and terminator. Figure 2.5 illustrates this principle for a known *E.coli histidine* operon⁵⁹.

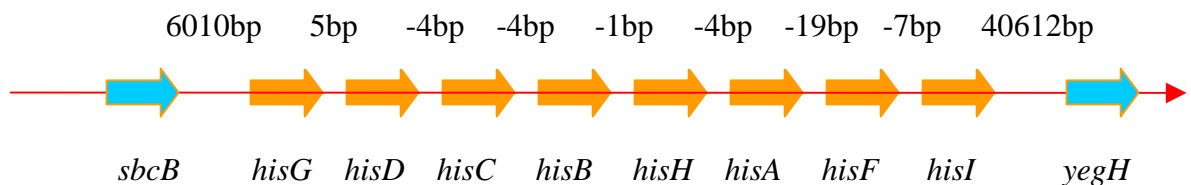


Figure 2.5. Example of an *E.coli* operon conforming to the expectations of the gene gap method. The inter-gene gaps in this operon range from -19 to 5bp, much smaller than the non-coding regions flanking this operon, 6010bp and 40612bp.

Specificity and Sensitivity of the Gene Gap Method

All intergene non-coding region lengths were calculated in *E.coli*. Figure 2.6A shows the distribution of inter-gene gaps over the range -50 to 100 base pairs. (A value of -50 indicates that the first gene's open reading frame overlaps the second one's by 50 base pairs). There is a concentration of values between -10 and 20 base pairs, with the peak value at -4. Most cases with a value of -4 have an overlap sequence of ATGA, with TGA acting as the stop codon for the upstream gene and ATG as the start codon for the downstream one. A gap of -1 is also common, and here the sequence bridging the two genes is usually T(A/G)ATG, with TAA/TGA as the stop codon for the upstream gene and ATG as the start codon for the downstream one. No gaps of length -3 or of multiples of -3 are observed. Such an arrangement would imply that the stop codon of the upstream codon be in frame in the downstream gene, obviously non-viable.

The method was benchmarked in a similar manner to the Gene Neighbor Method. The specificity and sensitivity were evaluated as a function of the maximum gap length allowed between members of an operon gene pair. Figures 2.6B and 2.6C show these results. As expected, the sensitivity increases (fewer false negatives) and

the specificity decreases (more false positives) as the allowed gap increases. A threshold of +25 provides the best compromise between specificity and sensitivity, with values of 95% and 68% respectively. This value was used for predictions of operon gene pairs. In *E.coli*, 1357 operon gene pairs are predicted. The number of predictions for all microbial genomes is listed in Table 2.3.

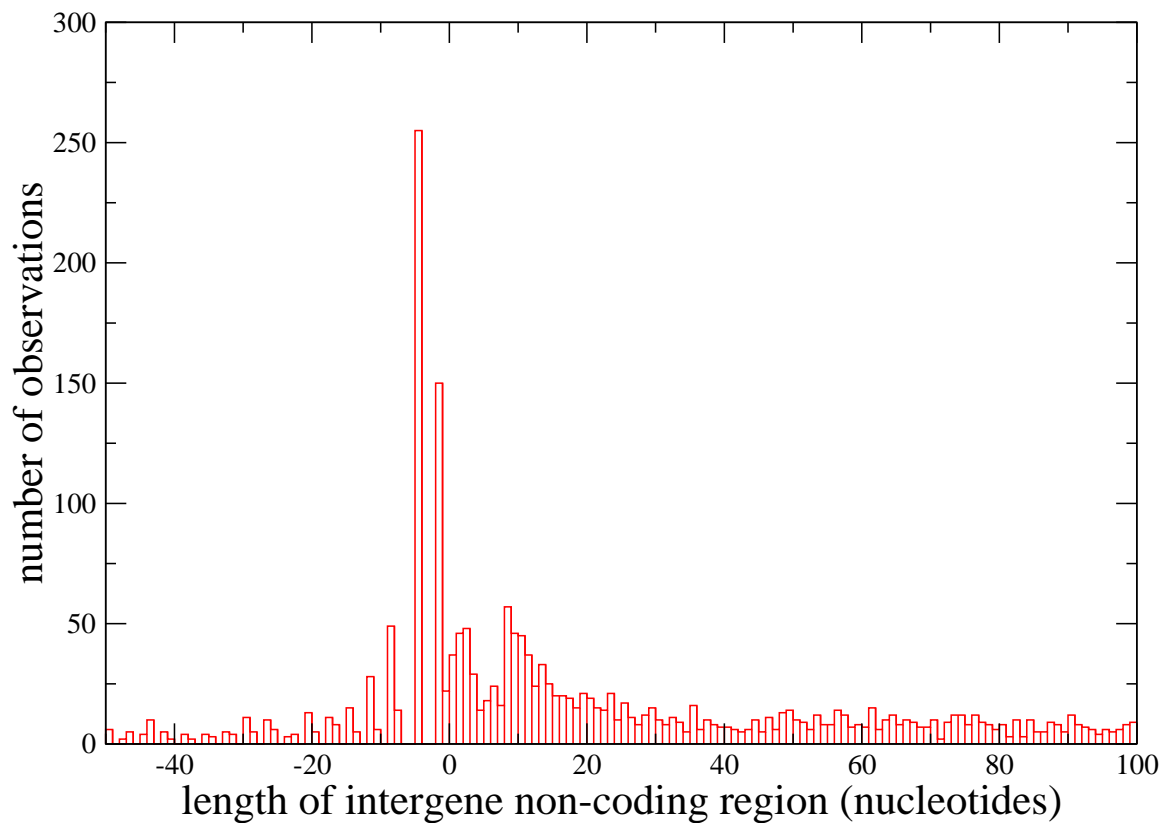


Figure 2.6A. Distribution of intergene gap lengths in *E.coli*. Negative values indicate overlapping genes. There is a concentration of gaps between -10 and 20 nucleotides, with the peak value at -4.

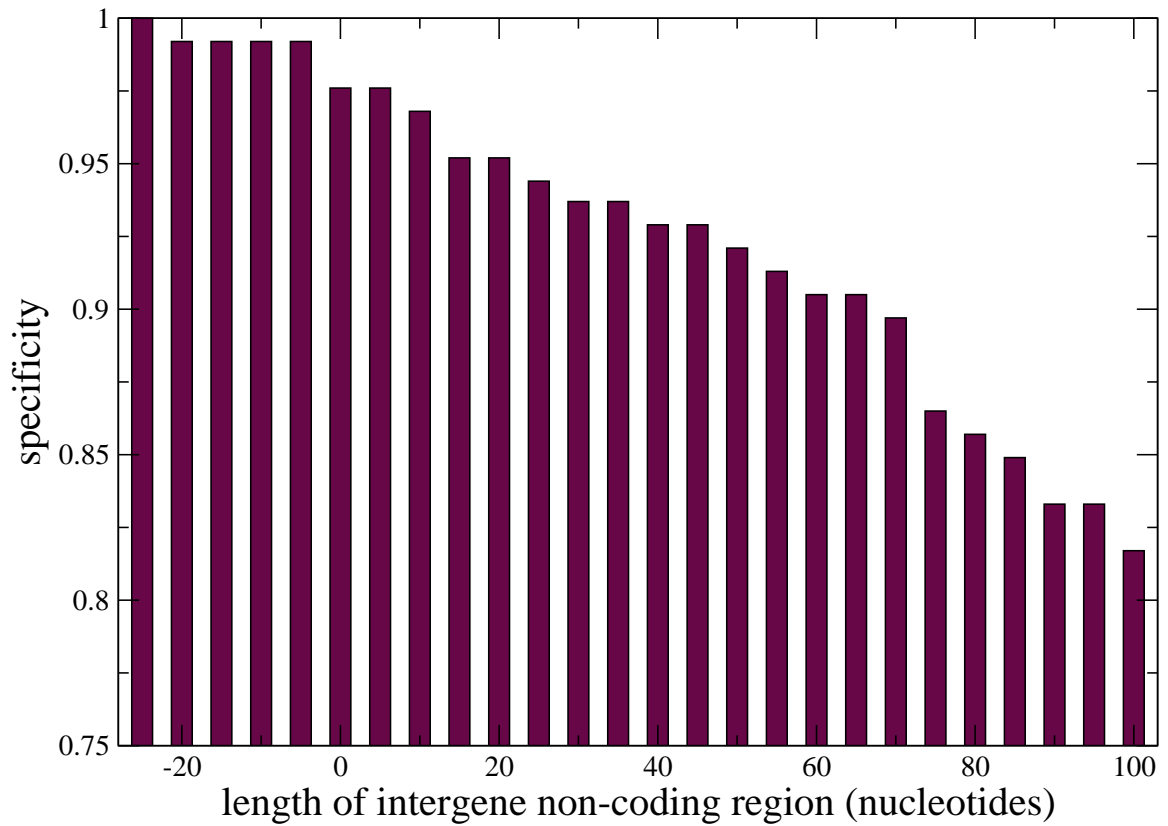


Figure 2.6B. Specificity of the Gene Gap method in *E.coli*, as a function of the maximum gap length allowed between members of an operon gene pair. The higher the threshold, the lower the specificity (i.e. there are more false positives). For a threshold of 25 nucleotides, the specificity is 95%.

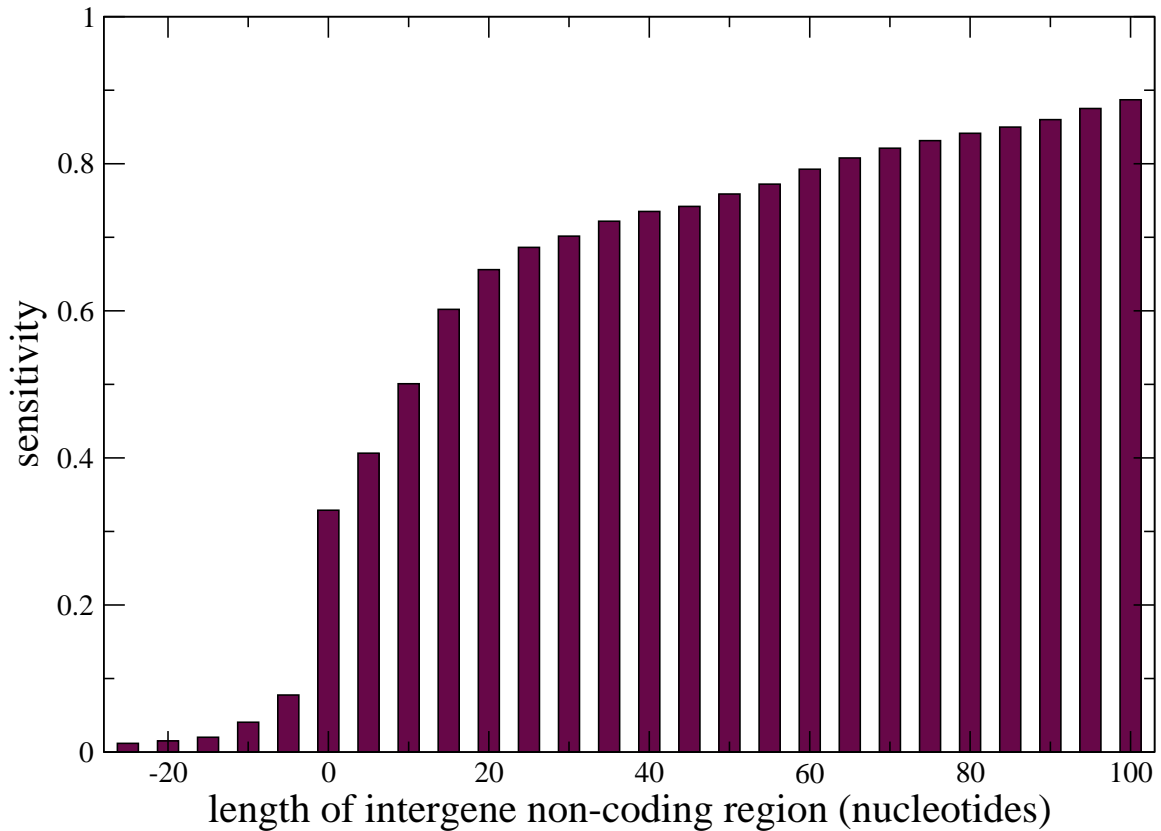


Figure 2.6C. Sensitivity of the Gene Gap method in *E.coli*, as a function of the maximum gap length allowed between members of an operon gene pair. The higher the threshold, the higher the sensitivity (i.e. there are fewer false negatives). For a threshold of 25 nucleotides, the sensitivity is 68%.

2.3.3 Methods Comparison

The conserved gene pair and intergene gap method provide independent sets of predictions. Thus, we would expect that consensus results, where the two methods agree, will be more reliable. Specificity should increase (fewer false positives), offset by lower sensitivity (fewer true positives). Figure 2.7 shows the results for the

common predictions, compared with those for the individual methods. As with the individual methods, specificity and sensitivity of the combined predictions were defined as:

$$Sp = TN/126$$

Where TN is the number of the 126 KEGG experimentally determined negatives identified by both methods. A value of 98% is obtained. That is, for cases where the two methods agree, there is only a 2% false positive rate.

$$Sn = TP/593$$

Where TP is the number of RegulonDB true positives identified by both methods. The two methods agree in 50% of these cases. Although this is a low coverage, as noted above, the reliability is high.

If increased sensitivity is desired, the methods can be used combined. One or both methods identify a true positive for 87% of cases, with a specificity of 94%.

Comparison of the two methods also provides a mechanism for independently evaluating the accuracy of the specificity and sensitivity. Given these values for two independent methods, we can calculate the expected values for the consensus method. For sensitivity, if $P_1(T)$ is the probability of identifying a true positive for method 1

and $P_2(T)$ is the corresponding for method 2, and these probabilities are independent, then the probability of both methods identifying the same true positive is:

$$P_{12}(T) = P_1(T) \cdot P_2(T) = 0.70 \cdot 0.68 = 0.49$$

The actual rate for the joint method is 0.50, a satisfactory result.

Expected specificity for the joint method is calculated as follows: The probability of method 1 producing a false positive as $P_1(F) = (1 - Sp_1)$, and similarly, $P_2(F) = (1 - Sp_2)$ for method 2, where Sp_1 and Sp_2 are the specificities of the two methods. Then the probability of both methods identifying the same false positive is

$$P_{12}(F) = P_1(F) \cdot P_2(F) = 0.07 \cdot 0.05 = 0.0035$$

yielding an expected specificity of 99.65%. The actual value is a little lower, at 98.5%, but still reasonably consistent with the benchmark results.

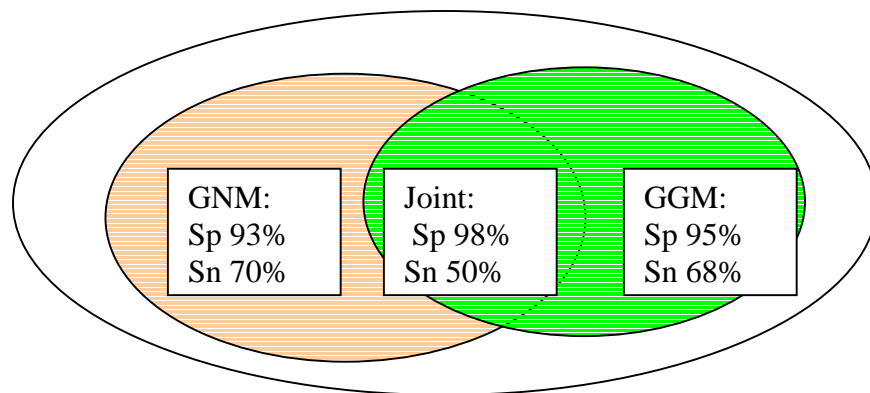


Figure 2.7. Comparison of operon gene pair predictions by the two methods, Specificity (Sp) and sensitivity (Sn) for the two methods are discussed as above, (GNM: gene neighbor method, GGM: gene gap method) Values for the overlap region are for cases where the predictions from the two methods agree. This yields a higher specificity (only 2% false positives) at the expense of lower coverage (sensitivity of 50%), quantitatively consistent with the sensitivity and specificity of the separate methods.

A further test of the operon prediction methods was obtained by predicting operons in the complete genome of *Saccharomyces cerevisiae*, a eukaryotic organism without operons. The gene neighbor method predicts only 32 operon gene pairs in this genome and gene gap method predicts only three.

2.3.4 Prediction of Protein Function

A primary motivation for predicting operons is to permit inference of the function of hypothetical protein. The idea ¹⁴ is that when an operon contains one or more genes with clear function assignment, that provide clues to the function of its other genes. Table 2.2 shows an example. An *H.influenzae* operon (predicted by both methods) that yields some functional insight.

Gene	Position in operon	Function Annotation	Conservation level to the previous / following gene	Inter-gene gap to the previous / following gene	Homolog in <i>E. coli</i>
HI1129	1	HYPOTHETICAL PROTEIN	0/13	201/30	YabB
HI1130	2	HYPOTHETICAL PROTEIN	13/15	30/2	YabC
HI1131	3	CELL DIVISION PROTEIN FTSL	15/15	2/12	FtsL
HI1132	4	PENICILLIN-BINDING PROTEIN 3	15/10	12/9	FtsI
HI1133	5	UDP-N-ACETYLMURAMOYLALANYL-D-GLUTAMATE--2,6-DIAMINOPIPELATE LIGASE	10/17	9/13	MurE
HI1134	6	UDP-N-ACETYLMURAMOYLALANYL-D-GLUTAMYL-2,6-DIAMINOPIPELATE--D-ALANYL-D-ALANYL LIGASE	17/11	13/-7	MurF
HI1135	7	PHOSPHO-N-ACETYLMURAMOYL-PENTAPEPTIDE-TRANSFERASE	11/13	-7/122	MraY
HI1136	8	UDP-N-ACETYLMURAMOYLALANINE--D-GLUTAMATE LIGASE	13/9	122/21	MurD
HI1137	9	CELL DIVISION PROTEIN FTSW	9/13	21/11	FtsW
HI1138	10	UDP-N-ACETYLGLUCOSAMINE--N-ACETYLMURAMYL-(PENTAPEPTIDE) PYROPHOSPHORYL-UNDECAPRENOL N-ACETYLGLUCOSAMINE TRANSFERASE	13/13	11/137	MurG
HI1139	11	UDP-N-ACETYLMURAMATE--ALANINE LIGASE	13/5	137/71	MurC
HI1140	12	D-ALANINE--D-ALANINE LIGASE	5/7	71/-1	DdlB
HI1141	13	CELL DIVISION PROTEIN	7/8	-1/18	FtsQ

FTSQ HOMOLOG					
HI1142	14	CELL DIVISION PROTEIN FTSA	8/16	18/83	FtsA
HI1143	15	CELL DIVISION PROTEIN FTSZ	16/4	83/38	FtsZ
HI1144	16	UDP-O- [3- HYDROXYMYRISTOYL] N- ACETYLGLUCOSAMINE DEACETYLASE	4/0	38/126	LpxC

Table 2.2. An example of a predicted operon in *H.influenzae*, yielding some functional insight. Genes are listed in the first column, and their positions in the operon are in the second column. The conservation levels of each gene with those before and after it are shown in the fourth column. The intergene gaps on either side of each gene are also shown, in the fifth column. Function annotations are taken from the Swiss-Prot database. (<http://www.expasy.ch/sprot/sprot-top.html>). The functions of the first and second genes in this operon are unknown. From the annotation of the other genes, it is clear that the operon is involved in cell wall structure formation and cell division. Thus it is likely that the two hypothetical proteins, HI1129 and HI1130, are also involved in these processes.

Table 2.3 lists the number of hypothetical proteins in each of the 42 genomes considered, and the total number of these that can potentially be partly annotated by the operon predictions. ‘Hypothetical’ are those proteins for which SwissProt release 40 annotation contains the words ‘hypothetical’, ‘unknown’ or ‘orf’. A hypothetical protein is considered partly annotatable if it is predicted to be in the same operon as one or more non-hypothetical proteins. Less well studied genomes tend to have a

larger fraction of hypothetical proteins, and so are less likely to have the required combination in an operon.

Genome	Number of genes	Hypo-thetical proteins	Hypo-thetical proteins annotated by GNM	Hypo-thetical proteins annotated by GGM	Total hypo-thetical proteins annotated	Percent of hypothetical proteins annotated in each genome
<i>Aeropyrum pernix</i>	2694	2065	31	89	107	5
<i>Aquifex aeolicus</i>	1522	663	17	254	260	39
<i>Archaeoglobus fulgidus</i>	2407	1439	59	319	334	23
<i>Bacillus halodurans</i>	4066	1925	79	319	355	18
<i>Bacillus subtilis</i>	4100	1912	52	163	193	10
<i>Borrelia burgdorferi</i>	1637	1099	20	143	149	14
<i>Buchnera sp. APS</i>	574	87	21	13	29	33
<i>Campylobacter jejuni</i>	1629	979	65	361	372	38
<i>Caulobacter</i>	3737	1810	55	374	387	21

<i>crescentus</i>						
<i>Chlamydia</i>	909	449	30	72	82	18
<i>muridarum</i>						
<i>Chlamydia</i>	893	295	7	51	55	19
<i>trachomatis</i>						
<i>Chlamydophila</i>	1052	435	11	72	77	18
<i>pneumoniae</i>						
<i>Chlamydophila</i>	1110	632	33	95	110	17
<i>pneumoniae AR39</i>						
<i>Chlamydophila</i>	1069	434	11	72	77	18
<i>pneumoniae J138</i>						
<i>Deinococcus</i>	3102	1884	43	283	299	16
<i>radiodurans</i>						
<i>Escherichia coli</i>	4289	1430	173	227	313	22
<i>Escherichia coli</i>	5361	1938	180	226	313	16
<i>O157:H7</i>						
<i>Haemophilus</i>	1709	707	103	165	209	30
<i>influenzae Rd</i>						
<i>Halobacterium sp.</i>	2605	1572	30	133	149	9
<i>NRC-1</i>						
<i>Helicobacter</i>	1566	691	16	236	238	34
<i>pylori 26695</i>						
<i>Helicobacter</i>	1490	618	15	242	244	39

<i>pylori J99</i>						
<i>Lactococcus lactis</i>	2266	825	21	127	144	17
<i>subsp. lactis</i>						
<i>Methanobacterium</i>	1869	499	6	146	147	29
<i>thermoautotrophic</i>						
<i>um</i>						
<i>Methanococcus</i>	1770	1054	29	198	215	20
<i>jannaschii</i>						
<i>Mycobacterium</i>	1605	987	37	117	130	13
<i>leprae</i>						
<i>Mycobacterium</i>	3918	2441	30	392	404	17
<i>tuberculosis</i>						
<i>Mycoplasma</i>	480	216	16	82	85	39
<i>genitalium</i>						
<i>Mycoplasma</i>	688	294	12	68	72	24
<i>pneumoniae</i>						
<i>Neisseria</i>	2025	995	54	145	178	18
<i>meningitidis</i>						
<i>Neisseria</i>	2032	708	39	174	196	28
<i>meningitidis Z2491</i>						
<i>Pasteurella</i>	2014	958	120	215	262	27
<i>multocida</i>						
<i>Pseudomonas</i>	5565	2545	145	298	375	15

<i>aeruginosa</i>						
<i>Pyrococcus abyssi</i>	1765	946	56	238	247	26
<i>Pyrococcus</i>	2064	1506	68	198	223	15
<i>horikoshii</i>						
<i>Rickettsia</i>	834	340	21	65	76	22
<i>prowazekii</i>						
<i>Synechocystis</i>	3169	1748	26	138	157	9
<i>PCC6803</i>						
<i>Thermoplasma</i>	1478	963	77	149	178	18
<i>acidophilum</i>						
<i>Thermotoga</i>	1846	975	61	404	424	43
<i>maritime</i>						
<i>Treponema</i>	1031	534	19	120	129	24
<i>pallidum</i>						
<i>Ureaplasma</i>	611	299	8	70	74	25
<i>urealyticum</i>						
<i>Vibrio cholerae</i>	3828	1846	128	312	374	20
<i>Xylella fastidiosa</i>	2831	1535	48	251	276	18

Table 2.3. The numbers of hypothetical proteins that can be partially annotated by the Gene Neighbor Method (GNM) and the Gene Gap Method (GGM), for 42 microbial genomes. Hypothetical proteins are those with Swiss-Prot (release 40) function annotation containing the word “hypothetical”, “unknown” or “orf”. A hypothetical

protein is considered partly annotatable if it exists in a predicted operon containing one or more non-hypothetical proteins.

2.3.5 Combination of Structure and Operon Information

Operon context may provide complementary function information to that from other sources. We have used operon information to supplement that obtained from structure in a structural genomics project focused on providing functional information for ‘hypothetical’ microbial proteins (s2f.umbi.umd.edu)^{60; 61}. Of 45 protein structures obtained, 11 (HI0065, HI0393, HI0442, HI0670, HI0817, HI1034, HI1333, HI1543, HI1679 in *Haemophilus influenzae*, YbgI, YqgF in *Escherichia coli*) are part of predicted operons where one or more proteins have assigned function (Table 2.4).

We find that in some cases, predicted operon context provides critical extra information for arriving at likely function, given a structure. This is particularly true when structure reveals that the protein is a member of a known superfamily, and so provides a rough indication of likely biochemical function. For example, the structure of HI1679 showed it to be a member of phosphatase superfamily⁶², and this protein is predicted to be in the same operon as HI1678. This latter protein was known to be arabinose-5-phosphate isomerase in the 3-Deoxy-D-manno-octulosonate (KDO) biosynthetic pathway⁶³. Together, these two pieces of information are sufficient, in principle, to pinpoint the function of HI1679 as 3-Deoxy-D-manno-octulosonate 8-

phosphate (KDO 8-P) phosphatase, the only phosphatase in KDO biosynthetic pathway⁶⁴. Even when the structure turns out not to belong to an already known superfamily, the operon information may still be useful. An example is the case of HI0442. The structure revealed a previously unknown and unusual fold, with a pronounced negative charge distribution on the surface⁶⁵. It occurs in a predicted operon with HI0443, a presumed ortholog to a known DNA repair enzyme. The unusual charge distribution and form of the fold, together with the operon context, led to the hypothesis that HI0442 is a double strand DNA mimic, sequestering DNA binding proteins, in a manner similar to DinI in *E. coli*⁶⁶.

Genome	Prediction Method and Signal	Gene name	Function Annotation
<i>Escherichia coli</i>	GGM (22, -7)	<i>ybgI</i> *	Hypothetical protein
		<i>ybgJ</i>	Putative carboxylase
		<i>ybgK</i>	Putative carboxylase
<i>Escherichia coli</i>	GNM (6,3,3)	<i>yqgF</i> *	Hypothetical protein
		<i>yqgE</i>	Hypothetical protein
		<i>gshB</i>	Glutathione synthetase
		<i>yggJ</i>	Hypothetical protein

<i>Haemophilus influenzae</i>	GGM (7,0,7,5) GNM (6,5,5,0)	HI0065*	Hypothetical protein
		HI0066	Probable N-acetylmuramoyl-L-alanine amidase amiB [Precursor]
		HI0067	DNA mismatch repair protein mutL
		HI0068	tRNA delta(2)- isopentenylpyrophosphate transferase
		HI0069	Glutamate-ammonia-ligase adenyltransferase
<i>Haemophilus influenzae</i>	GNM (6)	HI0393*	Hypothetical protein
		HI0394	Peptidyl-tRNA hydrolase
<i>Haemophilus influenzae</i>	GNM (3)	HI0442*	Hypothetical protein
		HI0443	Recombination DNA repair protein
<i>Haemophilus influenzae</i>	GGM (-4)	HI0669	Probable electron transporter required for biotin synthase activity
		HI0670*	Hypothetical protein
<i>Haemophilus influenzae</i>	GGM (11)	HI0816	Proline aminopeptidase
	GNM (3)	HI0817*	Hypothetical protein
<i>Haemophilus influenzae</i>	GGM (17)	HI1033	Phosphoserine phosphatase
		HI1034*	Hypothetical protein

<i>Haemophilus influenzae</i>	GNM (3)	HI1330	D-alanyl-D-alanine carboxypeptidase D-alanyl-D-alanine-endopeptidase
		HI1333*	Hypothetical protein
<i>Haemophilus influenzae</i>	GGM (-1)	HI1543*	Hypothetical protein
		HI1544	Putative NAD(P)H oxidoreductase
<i>Haemophilus influenzae</i>	GGM (5)	HI1678	Probable phosphosugar isomerase
	GNM (9)	HI1679*	Hypothetical protein

Table 2.4. Cases where operon context provides additional information for Hypothetical proteins whose structure has been obtained as part of a structural genomic project. Of 45 protein structures obtained, 11 (HI0065, HI0393, HI0442, HI0670, HI0817, HI1034, HI1333, HI1543, HI1679 in *Haemophilus influenzae*, *ybgI*, *yqgF* in *Escherichia coli*) are part of predicted operons where one or more proteins have assigned function. ‘GGM’ is the Gene Gap Method and ‘GNM’ is the Gene Neighbor Method. The value in brackets shows the signal supporting each predicted operon pair: for GGM, non-coding nucleotide intergene distance, and for GNM, the gene pair conservation level. Genes and their SwissProt annotation are shown in the last two columns. Structural genomics targets are indicated with *s.

2.3.6 Extent of Operon Conservation and implications for Pathway Consistency

Predicted operons may also be used to compare specific pathways in different organisms – if the two species have a set of equivalent proteins in an operon, for example, that associated with cell wall synthesis, it is likely that the pathways for performing that function are the same. (Other methods, such as phylogenetic profiles⁶⁷ may also be used for this purpose). Figure 2.8 shows the distribution of operon gene pair conservation across the 30 genome groups, for all *E.coli* operon genes. The majority of pairs can be detected for only three groups. Very few predicted pairs persist across even half of the genome groups. Figure 2.9 shows the distribution of size and conservation level for complete *E.coli* operons (the conservation level of an operon is defined as the smallest conservation level of any gene pair it contains). Most predicted operons are small (less than five genes) and the larger ones are very unlikely to be conserved in many genomes. The low level of conservation extends to groups of proteins known to form physical complexes. For example, succinate dehydrogenase, a critical enzyme in the TCA cycle of energy metabolism. In *E. coli*, four genes (*sdhC*, *D*, *A*, and *B*) in the *sdh* operon encode the four subunits of this enzyme. However, only in seven genomes (occurring in five genome groups) is this operon structure conserved. In two other genomes, homologs of all four genes are found, but they do not form an operon. In 25 genomes, there are no homologs of C and D, although A and B are still found, in some cases in an operon, and in some

cases not. No homologs of any of the genes can be found in the remaining eight genomes.

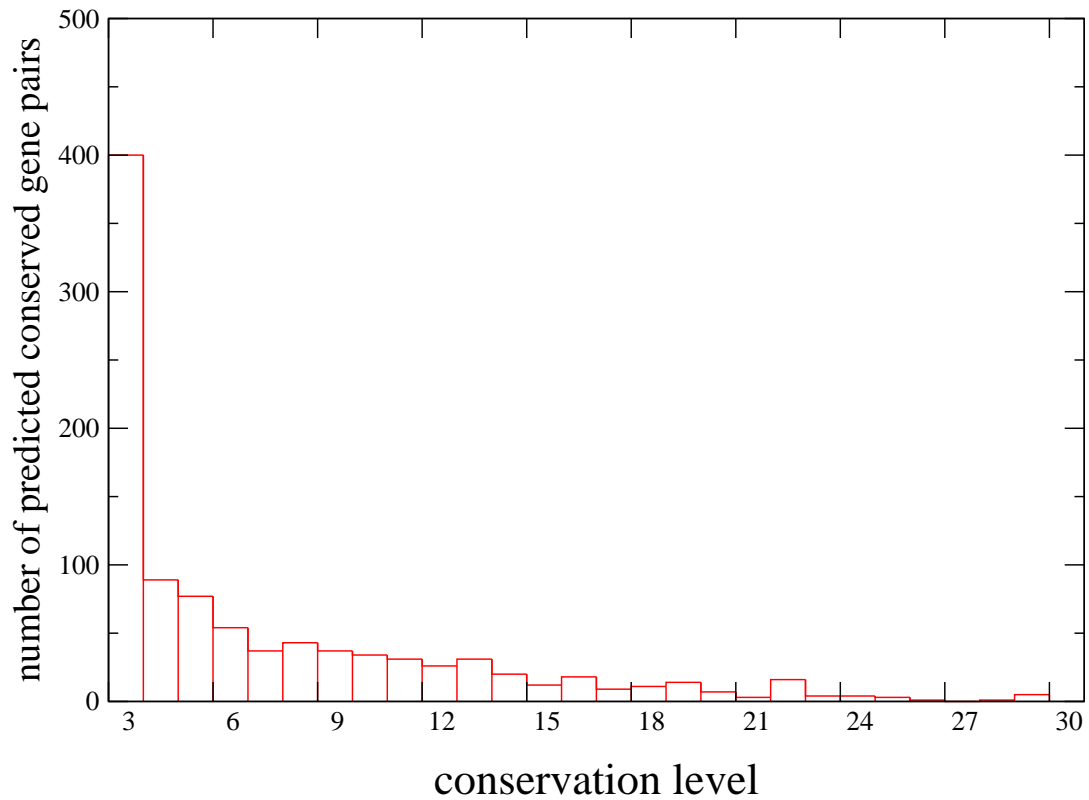


Figure 2.8. Histogram of the conservation level of predicted operon gene pairs in *E.coli*. Most operon gene pairs are conserved across only a few genome groups.

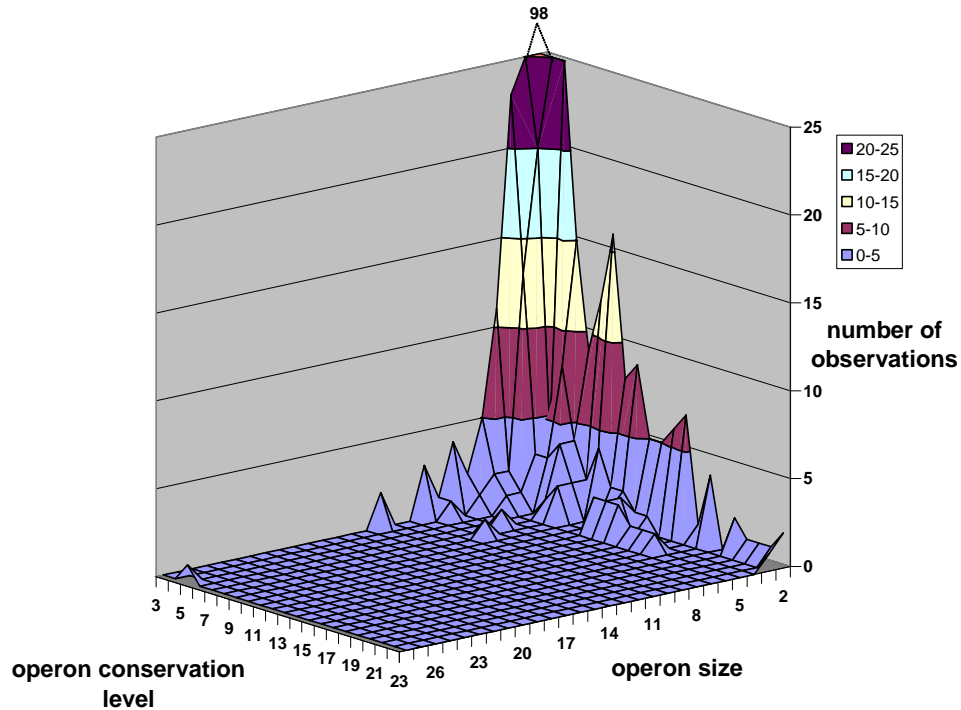


Figure 2.9. Number of predicted *E.coli* operons as a function of operon size (number of genes) and conservation level of the complete operon. Most of operons are small and conserved in only a few genome groups. There are rare cases of large size or conservation across many genome groups, but no instances with both characteristics. (For clarity, the top of the peak is truncated).

2.4 Conclusion and Discussion

Two operon detection strategies are presented in this paper, the Gene Neighbor Method and the Gene Gap Method. The Gene Neighbor method (GNM) utilizes the relatively high conservation of gene order in operons, compared with genes in

general. Two new concepts, the common neighbor fraction (CNF) and the genome group, are introduced to allow the quantification of gene order conservation. The Gene Gap Method (GGM) makes use of the relatively short gap between genes in operons compared with that otherwise found between adjacent genes.

Predictions made with both methods are benchmarked for *E.coli* operons using KEGG pathway data to assess specificity and RegulonDB *E.coli* operon data to assess sensitivity. For the GNM, requiring conservation of gene order in at least three genome groups, the specificity is 93% (7% false positives) and the sensitivity 70% (not detecting 30% of true operon pairs). For the GGM, selecting adjacent gene pairs with an inter-gene gap of less than 25 nucleotides, the specificity is 95% (5% false positives) and the sensitivity 68% (not detecting 32% of true operon pairs). A combination of the two methods boosts the specificity to 98%, at the cost of a reduced sensitivity of 50%. Reassuringly, predictions from the two methods agree to an extent very close to that expected on the basis of the individual benchmarks, providing independent confirmation that the specificity and sensitivity are reasonably accurate. A limitation of the benchmarking is that it can only be performed for *E.coli* operon pairs. It is not known to what extent gene order and intergene gap properties do or do not vary for different sub-groups of microbial organisms.

Varying the threshold parameters for either method allows a higher or lower specificity to be selected. For example, if an application requires a very low rate of false positives, a higher conservation level or smaller inter-gene gap may be used, or

only consensus predictions from the two methods accepted. Conversely, if hints of possible operon gene pairs are required, rather than high reliability, a conservation level of two or a larger inter-gene gap may be used.

The primary use of operon predictions is to acquire functional insight for unannotated proteins in microbial organisms. For the genomes considered, the fraction of unannotated proteins (according to Swiss-Prot release 40) varies between 15% in *Buchnera sp. APS* and 73% in *Pyrococcus horikoshii*. (other annotation sources, such as ⁶⁸ provide suggested functions for a higher fraction of orfs). Annotation based on operon prediction requires at least one member of an operon has assigned function. The fraction of proteins that can be partially annotated on this basis varies from a high of 43% for *Thermotoga maritime* to a low of 5% for *Aeropyrum pernix*. Well studied organisms provide more clues about function. However, functional information provided by this source alone is limited to the class of function that a protein is likely involved in, for example cell wall synthesis, and it is not possible to assign a molecular function.

Other function information, such as that provided by protein structure, can interact synergistically with that obtained from operon context. Our experience in a structural genomics project has been that many hypothetical proteins turn out to belong to known structural superfamilies (s2f.umbi.umd.edu) ^{60; 61}. The reason for this is that remote evolutionary relationships often cannot be detected at the sequence level, but are usually obvious at the structural level. Superfamily membership implies a low

resolution molecular function, such as ‘phosphatase’ or ‘GTPase’. The operon information provides equally imprecise clues to cellular role, such as ‘cell wall synthesis’ or ‘DNA repair’. As illustrated by the case of HI1679, the two different types of information can sometimes be combined to gain a more complete functional picture. Less commonly, even when the structure turns out not to be that of an already known superfamily or fold, the operon information may be useful, as in the case of HI0442.

A surprising finding is that although there is a useful degree of conservation of operon pairs, that rarely extends over a large number of genomes and rarely are whole operons containing three or more proteins conserved over more than a few genome groups. Failure to detect all sequence relatives may be a complicating factor in this analysis although for orthologs, the fraction of relationships detected is likely to be high. In some cases, particularly for larger operons, shuffling of gene order may make the conservation undetectable by the methods used here. Nevertheless, it is clear that there is a very high level of pathway variation within these organisms.

Other genome scale methods that provide some function information have been developed⁶⁹. These include *domain fusion* (some times called the ‘*Rossetta Stone*’ method)⁷⁰, which utilizes the fact that protein domains with related function may be found on the same polypeptide chain in some organisms, but as separate polypeptide chains in others; the *Phylogenetic profile* method^{67, 71}, which utilizes the fact that when two proteins from a target organism have homologs in the same subset of other

fully sequenced organisms, a related cellular function is suggested; and *Expression profile comparison*⁷², utilizing correlations between the expression profiles of sets of proteins. Our experience in the structural genomics project is that at present operon prediction is by far the most useful, although increasing amounts of data may in time make the others more competitive.

The predicted operon pairs and operons for the 42 microbial genomes are available at <http://moult.umbi.umd.edu/operons/>.

Chapter 3: Protein Family Clustering for Structural Genomics

3.1 Introduction

The ultimate goal of structural genomics is to provide structures for all biological proteins. Although there have been enormous improvements in experimental methods for determining structure⁷³, these still lag behind sequencing methods by orders of magnitude, in both cost and speed. As a result, currently, only about 1% of proteins with known sequence also have an experimentally known structure. Fortunately, it is not essential to experimentally determine the structure of every protein – evolutionally related proteins have similar structures^{74; 75}, and so comparative modeling methods can be used to obtain structure for any protein with a detectable evolutionary relationship to one with an experimental structure. This strategy has been widely accepted^{75; 76; 77; 78; 79;7}. The accuracy of comparative models depends on the closeness of the evolutionary relationships they are based on⁸⁰, and is never as high as that of a high quality X-ray structure. Nevertheless, these models are useful for many practical applications⁸¹.

The minimum number of experimental structures that will be needed in order to model all proteins using evolutionary relationships depends on the nature of protein sequence space. In particular, this number depends on how many families of evolutionarily associable proteins there are. The recent increase in fully sequenced genomes has made it possible to estimate this quantity more reliably than in the past. In this paper, we make use of knowledge of the full sequences for a set of 67 bacteria to obtain such an estimate.

No sequence based method is able to detect all evolutionary relationships: experimental structure determinations reveal previously undetectable relationships in many cases. Thus, all sequence based families are in some sense arbitrary, reflecting the effectiveness of current relationship detection algorithms rather than the number of independent evolutionary lines. From a structural genomics perspective, current methods are sufficiently powerful that they already represent very coarse grained sampling of structure space, so that models based on one experimental structure per family are probably at the limit of useful accuracy⁸². A single family will also often embrace a number of functions⁸³.

Clustering of proteins into families has long been used as a basis for extending function annotation, and so there is a history of algorithm development^{84 4; 85; 86; 87; 88; 89; 90; 91; 92; 93; 94}. Many of the family sets have been developed for specific purposes, and there is so far no universally accepted comprehensive source. For example, PfamA⁴, one of the best established sets, uses sensitive methods to detect remote

evolutionary relationships, and is hand curated, providing high reliability. As a consequence, coverage is incomplete.

We have developed an automated family classification scheme, applicable to estimation of the number of experimental structures that will be needed for structural genomics. There are three main steps: identification of evolutionary relationships, parsing of the full proteins into probable structural domains, and clustering into families. Conventional PSI-BLAST searches are used to detect sequence relationships within a set of 67 fully sequenced bacterial genomes. Lists of relationships are subdivided on the basis of a protein domain identification method. Lists are then merged into families with a multi-linkage clustering procedure. Although a relatively standard sequence search method is used, benchmarking with SCOP structural superfamilies^{3;2} shows a slightly higher sensitivity than previously reported methods including profile and profile-profile methods <http://supfam.mrc-lmb.cam.ac.uk/PRC/>^{95 96 97}. We attribute this to the robust clustering step and reasonably effective parsing into domains.

There have been a number of studies of the number of protein families in biology. Estimates vary from 1000 to 30,000^{7; 76; 79; 98; 99; 100; 101; 102; 103; 104}. As more genome sequences are completed, it becomes possible to improve the reliability of the estimate. Our study, focusing on recently available complete genome sequences, leads to an estimate for the prokaryotes that is substantially higher than previous ones: Clustering 178,310 sequences from 67 microbial genomes already generates

31,874 families. A recent study of five fully sequenced eukaryotic genomes ¹⁰⁵ has also led to a much larger number than previously suggested, 45,000 protein families. A more relevant quantity for structural genomics is the number of detectable families there will be in future. We have developed a method of estimating growth in the number of families, and find there will be about 250,000 families when 1000 genomes are sequenced. Apparent singletons (proteins with no detectable relatives ⁶) are the fastest growing category.

At first glance, these increased estimates are discouraging for the structural genomics goal of obtaining structures for all domains. However, because most sequences are in relatively large families, we estimate that it will still be possible to have coverage for 70%~80% of domains within the next decade.

3.2 Methods

Protein sequences

All identified protein sequences in 140 genomes were retrieved from Genbank (<http://www.ncbi.nlm.nih.gov/Genomes/index.html>). 67 of these were used for building the family estimate model, and the rest were reserved for testing the projections of the model. All downloaded and generated information were stored in a MySQL relational database running on a Linux server.

Genome	Number of proteins
<i>Aeropyrum pernix</i>	2694
<i>Agrobacterium tumefaciens str. C58 (Dupont)</i>	5402
<i>Aquifex aeolicus</i>	1553
<i>Archaeoglobus fulgidus</i>	2407
<i>Bacillus halodurans</i>	4066
<i>Bacillus subtilis</i>	4100
<i>Borrelia burgdorferi</i>	1637
<i>Brucella melitensis</i>	3198
<i>Buchnera sp. APS</i>	574
<i>Campylobacter jejuni</i>	1629
<i>Caulobacter crescentus</i>	3737
<i>Chlamydia muridarum</i>	916
<i>Chlamydophila pneumoniae AR39</i>	1110
<i>Chlamydophila pneumoniae CWL029</i>	1052
<i>Chlamydophila pneumoniae J138</i>	1069
<i>Clostridium acetobutylicum</i>	3672
<i>Clostridium perfringens</i>	2723
<i>Corynebacterium glutamicum</i>	3040
<i>Deinococcus radiodurans</i>	3102
<i>Escherichia coli</i>	4289
<i>Escherichia coli O157:H7</i>	5361
<i>Haemophilus influenzae Rd</i>	1709
<i>Halobacterium sp. NRC-1</i>	2605
<i>Helicobacter pylori 26695</i>	1566
<i>Helicobacter pylori J99</i>	1490
<i>Lactococcus lactis subsp. lactis</i>	2266
<i>Listeria innocua</i>	3043
<i>Listeria monocytogenes EGD-e</i>	2846
<i>Mesorhizobium loti</i>	7275
<i>Methanobacterium thermoautotrophicum</i>	1869
<i>Methanococcus jannaschii</i>	1770
<i>Mycobacterium leprae</i>	1605
<i>Mycobacterium tuberculosis</i>	3869
<i>Mycobacterium tuberculosis CDC1551</i>	4187
<i>Mycoplasma genitalium</i>	480
<i>Mycoplasma pneumoniae</i>	688
<i>Mycoplasma pulmonis</i>	782
<i>Neisseria meningitidis</i>	2025

<i>Neisseria meningitidis</i> Z2491	2032
<i>Nostoc</i> sp. PCC 7120	6129
<i>Pasteurella multocida</i>	2014
<i>Pseudomonas aeruginosa</i>	5565
<i>Pyrobaculum aerophilum</i>	2605
<i>Pyrococcus abyssi</i>	1765
<i>Pyrococcus horikoshii</i>	2064
<i>Ralstonia solanacearum</i>	5116
<i>Rhizobium</i> sp. NGR234	416
<i>Rickettsia conorii</i>	1374
<i>Rickettsia prowazekii</i>	834
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Typhi	4749
<i>Salmonella typhimurium</i> LT2	4553
<i>Sinorhizobium meliloti</i>	6205
<i>Staphylococcus aureus</i> subsp. <i>aureus</i> Mu50	2748
<i>Staphylococcus aureus</i> subsp. <i>aureus</i> N315	2624
<i>Streptococcus pneumoniae</i>	2094
<i>Streptococcus pyogenes</i>	1696
<i>Sulfolobus solfataricus</i>	2977
<i>Sulfolobus tokodaii</i>	2826
<i>Synechocystis</i> PCC6803	3169
<i>Thermoplasma acidophilum</i>	1478
<i>Thermoplasma volcanium</i>	1526
<i>Thermotoga maritima</i>	1846
<i>Treponema pallidum</i>	1031
<i>Ureaplasma urealyticum</i>	611
<i>Vibrio cholerae</i>	3828
<i>Xylella fastidiosa</i>	2831
<i>Yersinia pestis</i>	4039

Table 3.1. 67 fully sequenced microbial genomes used for protein family construction, and the number of proteins in each genome. 12 are archaeal, and 55 are bacterial. In total there are 178,310 protein sequences.

Genome	Number of proteins
<i>Agrobacterium tumefaciens</i> str. C58 (Cereon)	5299
<i>Bacillus anthracis</i> str. Ames	5311
<i>Bacillus cereus</i> ATCC 14579	5255
<i>Bacteroides thetaiotaomicron</i> VPI-5482	4778
<i>Bifidobacterium longum</i> NCC2705	1729

<i>Bordetella bronchiseptica</i>	4994
<i>Bordetella parapertussis</i>	4185
<i>Bordetella pertussis</i>	3446
<i>Bradyrhizobium japonicum</i> USDA 110	8317
<i>Brucella suis</i> 1330	3264
<i>Buchnera aphidicola</i> str. Bp (<i>Baizongiapistaciae</i>)	504
<i>Buchnera aphidicola</i> str. Sg (<i>Schizaphisgraminum</i>)	546
<i>Candidatus Blochmannia floridanus</i>	583
<i>Chlamydia trachomatis</i>	893
<i>Chlamydophila caviae</i> GPIC	1005
<i>Chlamydophila pneumoniae</i> TW-183	1113
<i>Chlorobium tepidum</i> TLS	2252
<i>Chromobacterium violaceum</i> ATCC 12472	4407
<i>Clostridium tetani</i> E88	2373
<i>Corynebacterium efficiens</i> YS-314	2950
<i>Coxiella burnetii</i> RSA 493	2009
<i>Enterococcus faecalis</i> V583	3113
<i>Escherichia coli</i> CFT073	5379
<i>Escherichia coli</i> O157H7_EDL933	5349
<i>Fusobacterium nucleatum</i> subsp. <i>nucleatum</i> ATCC25586	2067
<i>Haemophilus ducreyi</i> 35000HP	1717
<i>Helicobacter hepaticus</i> ATCC 51449	1875
<i>Lactobacillus plantarum</i> WCFS1	3009
<i>Leptospira interrogans</i> serovar <i>lai</i> str. 56601	4727
<i>Methanopyrus kandleri</i> AV19	1687
<i>Methanosarcina acetivorans</i> C2A	4540
<i>Methanosarcina mazei</i> Goel	3371
<i>Mycobacterium bovis</i> subsp. <i>bovis</i> AF2122/97	3920
<i>Mycoplasma gallisepticum</i> R	726
<i>Mycoplasma penetrans</i>	1037
<i>Nitrosomonas europaea</i> ATCC 19718	2461
<i>Oceanobacillus iheyensis</i> HTE831	3500
<i>Pirellula</i> sp.	7325
<i>Porphyromonas gingivalis</i> W83	1909
<i>Prochlorococcus marinus</i> str. MIT 9313	2265
<i>Prochlorococcus marinus</i> subsp. <i>marinus</i> str. CCMP137	1882
<i>Prochlorococcus marinus</i> subsp. <i>pastoris</i> str. CCMP13	1712
<i>Pseudomonas putida</i> KT2440	5350
<i>Pseudomonas syringae</i> pv. <i>tomato</i> str. DC3000	5471
<i>Pyrococcus furiosus</i> DSM 3638	2065
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar <i>Typhi</i> T	4323
<i>Shewanella oneidensis</i> MR-1	4472
<i>Shigella flexneri</i> 2a str. 2457T	4068
<i>Shigella flexneri</i> 2a str. 301	4180
<i>Staphylococcus aureus</i> subsp. <i>aureus</i> MW2	2632

<i>Staphylococcus epidermidis</i> ATCC 12228	2419
<i>Streptococcus agalactiae</i> 2603V/R	2124
<i>Streptococcus agalactiae</i> NEM316	2094
<i>Streptococcus mutans</i> UA159	1960
<i>Streptococcus pneumoniae</i> R6	2043
<i>Streptococcus pyogenes</i> MGAS315	1865
<i>Streptococcus pyogenes</i> MGAS8232	1845
<i>Streptococcus pyogenes</i> SSI-1	1861
<i>Streptomyces avermitilis</i> MA-4680	7575
<i>Streptomyces coelicolor</i> A3(2)	8154
<i>Synechococcus</i> sp. WH 8102	2517
<i>Thermoanaerobacter tengcongensis</i>	2588
<i>Thermosynechococcus elongatus</i> BP-1	2475
<i>Tropheryma whipplei</i> str. Twist	808
<i>Tropheryma whipplei</i> TW08/27	783
<i>Vibrio parahaemolyticus</i> RIMD 2210633	4832
<i>Vibrio vulnificus</i> CMCP6	4537
<i>Wigglesworthia glossinidia</i> endosymbiont of <i>Glossina</i>	611
<i>Wolinella succinogenes</i>	503
<i>Xanthomonas axonopodis</i> pv. <i>citri</i> str. 306	4312
<i>Xanthomonas campestris</i> pv. <i>campestris</i> str. ATCC339	4181
<i>Xylella fastidiosa</i> Temecula1	2036
<i>Yersinia pestis</i> KIM	4090

Table 3.2. 73 recently sequenced microbial genomes used to test the family growth projections. In all, the 140 genomes code for 405,709 proteins.

Generation of Homolog Lists

For each protein, a six round PSI-BLAST search^{29; 30} was performed against the set of all other sequences in the genome set. Low complexity regions were omitted, using the default SEG¹⁰⁶ option. Homologs with an E-value 10^{-4} or lower to the search sequence were collected, creating a homolog list for each protein.

Domain Parsing

Each homolog list was examined for domain structures, as described below. A number of domain parsing methods have been developed^{87; 92 107}. In the present work, domain boundaries are identified based on the location of indels in the PSI-BLAST sequence alignment. Indel locations are found by counting the number of sequences with an amino acid at each position in the alignment. Figure 3.1 shows an example.

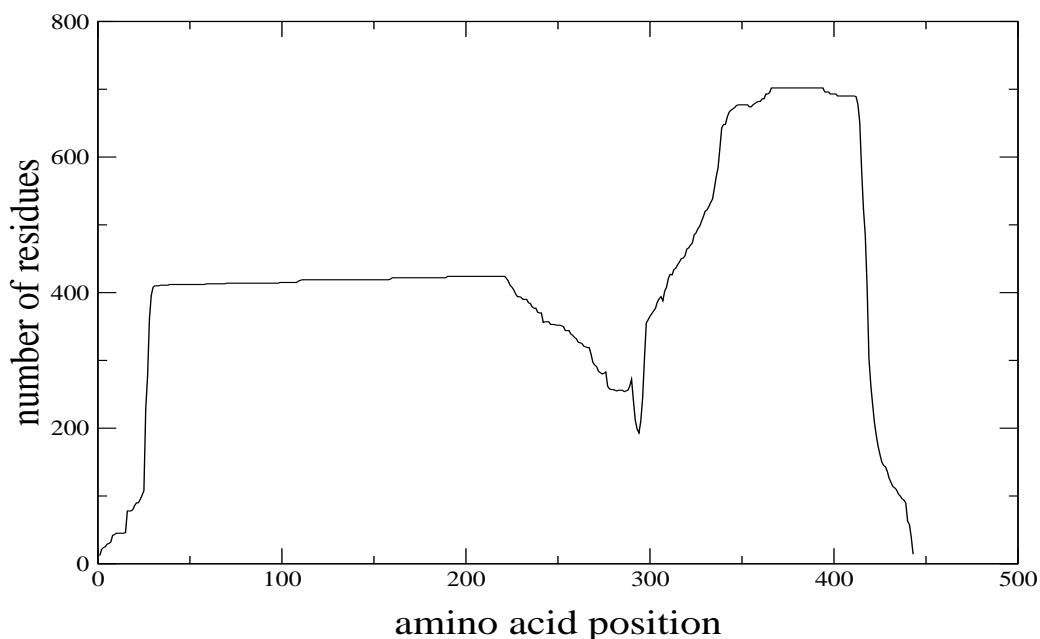


Figure 3.1. An example of domain parsing, for the multiple sequence alignment of *E. coli* ARG (swissprot ID P08205, Amino-acid acetyltransferase). The domain splitting algorithm produces two domains, residues 20-294, and 295- 443. Pfam and InterPro also split this protein into two domains. Domain 1 (26-269) belongs to Pfam PF00696, an amino acid kinase family, and domain 2 (338-414) belongs to PF00583, an acetyltransferase family.

Domain boundaries are defined as positions in the multiple alignment where there are relatively deep minima in the number of sequences with residues. The detailed procedure is as follows:

1. Calculate the slope of the alignment count for each position in the alignment.
2. Find all the turning points (positions where the sign of the slope changes).
3. Discard the trough points which make a domain too short to be viable (less than 40 residues between turning points).
4. Discard the trough points where a trough is not significantly lower than the surroundings (Trough height more than 60% of the peaks on either side).
5. Divide the proteins in the homolog list into domains by cutting at each remaining trough point, to create homolog domain lists.

As described later, comparison of the results of this procedure with a set of PfamA domains in 50,000 randomly chosen Pfam sequences shows it is very conservative: 96% of single domain PfamA proteins are predicted as such, but only 24% of PfamA two domains proteins are predicted correctly. Other domain parsers have similar accuracy, but adopt a different balance of false positives and false negatives ¹⁰⁵.

While this and other parsers are far from satisfactory, domain parsing does improve the quality of the families.

Merging of Domain Lists

Domain lists are highly redundant in that many domains appear in multiple lists. A key step is merging of the lists to form non-redundant domain-based protein families. Merging also increases the range of evolutionary relationships that are clustered: A PSI-BLAST search starting from protein A may find a relative B, but not relative C. On the other hand, PSI-BLAST started from protein B may have found relative C, but not relative A. Merging of the A and B hit lists places A, B and C in one family.

The simplest clustering procedure is to iteratively merge all pairs of lists that contain at least one common domain, and then eliminate redundancies from the merged sets. Notoriously, this single linkage procedure leads to over-clustering, even when the false positive rate for inclusion of a domain in a single list is small. A number of strategies have been suggested for overcoming this problem^{93; 108}. We have developed a variable linkage procedure. Short domain lists are merged on the basis of a single common entry. The longer the lists, the more common entries are required.

Merging proceeds by selecting a first list, comparing it to all others, combining where the merge rules are satisfied, then picking a next so far unconsidered list, and so on,

until all lists have been considered. The process is repeated a maximum of three times.

Further Domain Boundary Checking

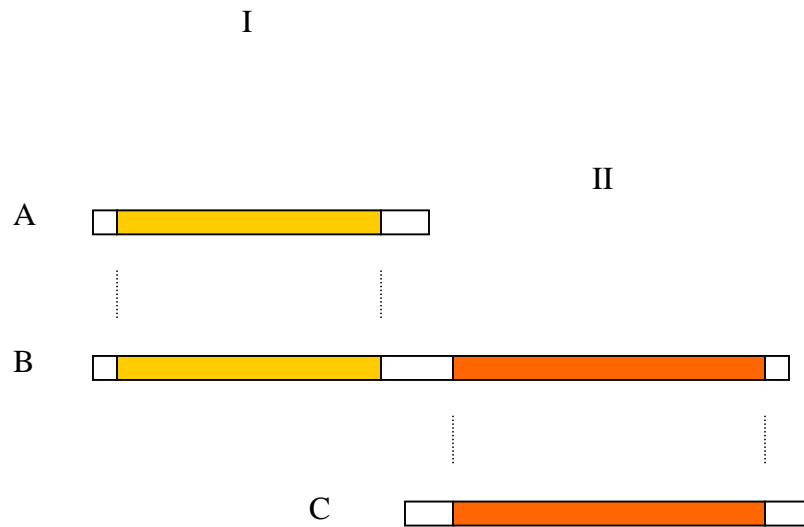


Figure 3.2. Domain Merging Check.

Incomplete domain parsing can occasionally lead to the merging of proteins that have no significant alignment. This is illustrated in figure 3.2. An unparsed two domain protein (B) in list I has a region of alignment with protein C in a second list, II. Sequence C shares no significant relationship with the primary domain in list I, but will be merged into that list. To reduce this effect, each candidate sequence in list II is

checked for alignment overlap with the first sequence in list I. List II entries with less than a 40 residue overlap are not merged.

Selection of Parameter Values

Results are dependent on a number of parameters. Parameters were optimized by building a set of families for proteins with domains in PfamA, varying parameter values to maximize the agreement between the generated and Pfam families, as others have done ¹⁰⁹. A set of 50,000 full length SwissProt protein sequences including all the domains present in a 721 family subset of PfamA version 9.0 was used. Use of full length protein sequences allows the domain parsing procedures to be tested.

Each generated family was compared with all the PfamA families, and the most similar one (most common sequences) was considered the best match. Two measures are used to assess the quality of the built families:

False negative fraction $FN = (P - O) / P$

False positive fraction $FP = (M - O) / M$

where P is the Pfam family size, M the generated family size, and O the common sequences between the two. FN is the false negative rate for a generated family – the fraction of correct domains omitted. FP is the false positive rate for a family – the fraction of incorrect domains included.

These ratios were determined for a range of PSI-BLAST conditions, with and without domain checks, and with different linkage rules, in order to optimize the procedure. Details are given in the 'Results' section.

The final choice of parameters was up to 6 rounds PSI-BLAST with an E-score threshold of 10^{-4} , and a maximum of three rounds of merging. Lists are merged into a family using the following merging rules:

For lists with four or fewer members: at least one common entry required for merging.

For lists with 5 to 10 members: at least two common entries.

For lists with more than 10 members: at least 40% common entries.

Evaluation of Domain Family Construction

Effectiveness of the family building procedure was assessed in terms of its ability to pair all members of SCOP superfamilies, and not to pair domains in different folds. SCOP (Structural Classification of Proteins, <http://scop.mrc-lmb.cam.ac.uk/scop/>) is a hierarchical organization of proteins based on evolutionary and structural relationships^{2; 3}. Since structural similarity provides a much more sensitive test of evolutionary relationships between proteins than does sequence, SCOP has been widely used as a benchmark for evaluating sequence alignment, clustering, and evolutionary relationship detection methods^{108; 110}. We have used SCOP40 (no

sequence relationships higher than 40% identity) version 1.63, which contains 5226 domains, 1224 superfamilies, and 760 folds.

The 5226 domains were clustered into families, as described above: PSI-BLAST was run for each domain against the NR sequence database, augmented with the SCOP domain sequences. No domain parsing was performed, as SCOP is already domain based. PSI-BLAST generated homolog lists were merged using the linkage rules, to form a set of generated families.

The set of generated families was compared with the SCOP superfamilies in terms of all the possible pairwise relationships between domains. Any pair of domains found both in a generated family and a SCOP superfamily is considered a true positive. A pair of domains presented in a generated family set, but not assigned the same SCOP fold is considered a false positive, as it is unlikely to represent a homology relationship. SCOP40 version 1.63 was used, with 50,374 pairs of domains within the same superfamily, and more than 600,000 pairs of domains with each member in a different fold. True positives detected as a function of the false positives incurred were plotted in a ROC curve. A 1% false positive to true positive ratio was chosen as an overall measure of quality, as used by others^{111; 112; 113; 114}.

For comparison, several other alignment and family clustering methods were also tested using the same set of SCOP domains. These are BLAST, PSIBLAST, SAM-T99 (HMM) (<http://www.soe.ucsc.edu/research/compbio/sam2src/>)¹¹⁵ and PRC (<http://supfam.mrc-lmb.cam.ac.uk/PRC/>) (a profile to profile method). Software was downloaded from the authors' web sites.

Programs and parameters used for SAM-99 were:

1. target99 –seed [sequence fasta file] –out [output file] –db [nr+scop40] –iter 4;
2. fw0.7 [sequence.a2m file] [sequence.mod file];
3. hmmscore [sequence name] –i [sequence.mod file]–sw 2 –db [scop40]

Programs and parameters used for PRC were:

Prc –Emax 10 [sequence.mod file] [mod library] [sequence name]

Transmembrane protein determination

Proteins with one or more transmembrane helical segments were identified using TMHMM (<http://www.cbs.dtu.dk/services/TMHMM/>)¹¹⁶.

Structure coverage determination

Domain families with known structures were identified as follows. A sequence profile (Position Specific Scoring Matrix, PSSM) was obtained from the multiple sequence alignment of a protein family, using blastpgp²⁹. Each protein sequence in the PDB (June 15, 2003 release) was run against the set of family sequence profiles, using RPS-BLAST (Reverse Position Specific BLAST)¹¹⁷. Any profile to sequence comparison with an E value of 10^{-2} or lower was considered to represent a family which could be modeled based on the corresponding structural template. Such families were considered to be structurally covered.

3.3 Results

3.3.1 Protein Family Clustering

Domain-based protein family set

Following the clustering procedure described in Methods, 178,310 sequences from 67 sequenced prokaryotic genomes were parsed to 249,574 domain sequences and then clustered into 31,874 sequence families. Figure 3.3A shows the distribution of family sizes. Small families predominate. There are 20,992 singletons (families with only one member), about 2/3 of the total, and 4810 doubletons (family size 2). At the other end of the spectrum, there are only 263 families larger than 100.

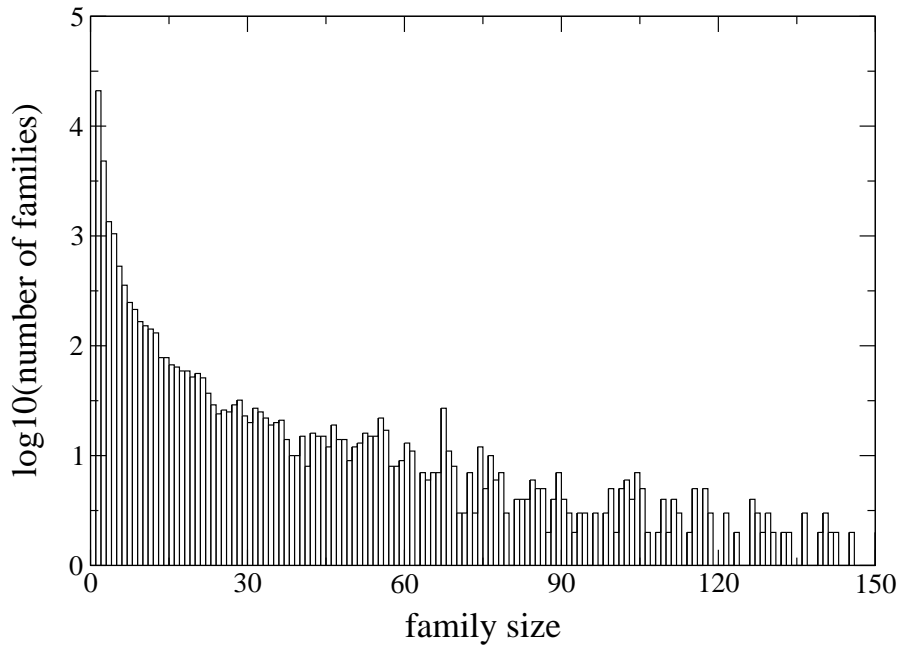


Figure 3.3A. Distribution of Domain Family Size. Note the log scale. There is an approximately power law relationship between the number of families and family size: 20,992 of the 31,874 families have only a single member, while only 263 families are larger than 100.

From the point of view of structural genomics, this result is discouraging: even this small number of genomes would require over 30,000 experimental structure determinations in order to provide templates for complete modeling. However, consideration of the high fraction of proteins in the larger families leads to a different view. Figure 3.3B compares the number of families of size 1, 2 and larger with the total number of domains those categories contain. Although about 2/3 of the families are singletons, they represent only 8% of the domains. Families of size 3 and larger contain 88% of the domains, and there are only just over 6000 of those. Thus, 88%

structural coverage of these 67 genomes would be provided by about 6000 experimental structure determinations.

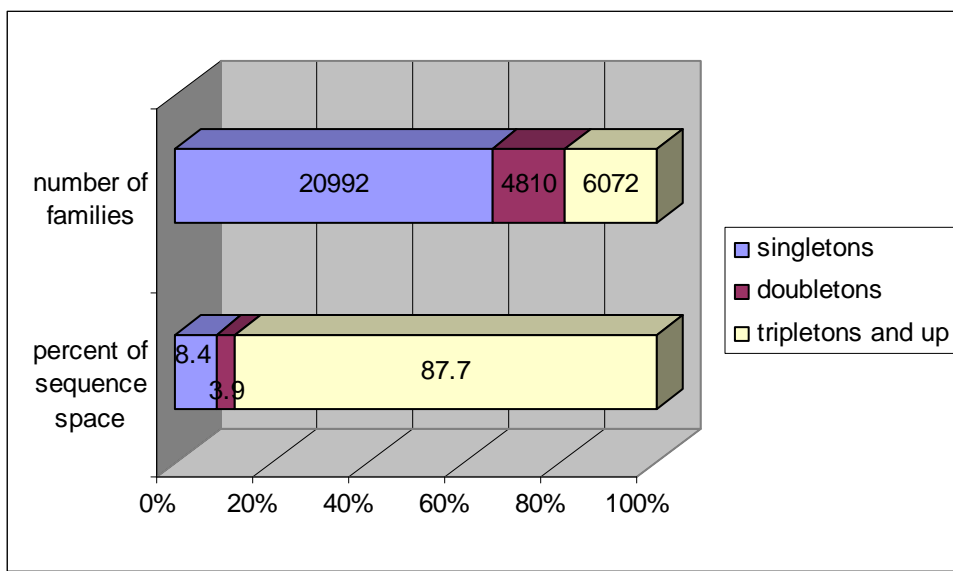


Figure 3.3B. Number of singletons (family size 1), doubletons (family size 2), tripletons and larger (top bar), and the percentage of sequence space covered by each of the three categories. Although there are 20,992 singletons and 4,810 doubletons, these two categories only represent about 12% of sequence space. The 6,072 larger families make up the rest.

Optimization of Protein Family Construction

As discussed in Methods, parameters for protein family construction were optimized by comparison of generated families with those in PfamA ⁴.

Families were built for 50,000 full length protein sequences covering 721 PfamA (release 9) families. The full sequences were clustered into new families, and each such family was best matched to a PfamA family. Each generated family was compared with the corresponding PfamA one using the false positive (FP) and false negative (FN) fractions. The smaller these values, the better the family building procedure.

Table 3.3 shows the level of agreement between the generated and PfamA families as a function of the E-value threshold for accepting PSI-BLAST relationships. Families are obviously over-clustered with a cutoff of 10^{-2} , judging by the high level of false positives (FP). A threshold of 10^{-4} produces many fewer false positives than 10^{-2} and a lower number of false negatives than 10^{-6} , so was chosen as the final value. (Final values of the other parameters were used for these tests.)

PSI-BLAST E-score Threshold	Number of families	FN(false negatives)	FP(false positives)	Number of Identical families
10^{-2}	569	0.087	0.325	204
10^{-4}	744	0.079	0.180	276
10^{-6}	788	0.087	0.176	273

Table 3.3. Agreement between generated and PfamA families, as a function of the PSI-BLAST E score threshold.

Table 3.4 shows the agreement between the generated and PfamA families as a function of the linking procedure, domain parsing and checking, and the number of merging rounds. A maximum of six rounds PSI-BLAST with an E score threshold of 10^{-4} was used. Three rounds of single linkage clustering with no domain processing dramatically over-clusters compared with PfamA, compressing the sequences into 285 families, as opposed to the ideal 721, with a false positive rate of 79%. Domain parsing increases the number of families to 427, at the expense of a minor increase in false negatives, from 6.9% to 8.0%. Domain checking produces a further minor improvement.

Introduction of the family size dependent linkage scheme further improves agreement with PfamA. Three rounds of merging generate 785 families with a false positive rate of 17.9%. Merging for 5 rounds slightly increases the false positive rate to 19.7%.

On the basis of these tests, the final protocol adopted was six rounds of all against all PSI-BLAST using a 10^{-4} threshold, followed by three rounds of hierarchical linkage. These conditions produce 785 families, of which 278 are identical to the corresponding PfamA ones, with an average false positive rate of 17.9% and a false negative rate of 7.4%. PfamA families are assembled using sensitive sequence methods and are hand curated to reduce false negatives, so that a good clustering

method should have a low false negative rate, as seen here. The higher false positive rate may partly reflect the fact that Pfam does not cluster some real relationships.

Clustering method	Number of generated families	FN	FP	Number of Identical families
Single linkage w/o domain splitting or domain check	285	0.069	0.792	154
Single linkage w/o domain check	427	0.080	0.415	244
Single linkage w/ domain check	480	0.079	0.377	255
Hierarchical merging, 3 rounds	785	0.074	0.179	278
Hierarchical merging, 5 rounds	744	0.079	0.197	276

Table 3.4. Agreement between generated and PfamA families as a function of linkage protocol, domain parsing and checking, and the number of rounds of merging. Domain parsing, domain checking, and hierarchical linkage all improve the quality of the generated families. On the basis of these results, a protocol of three rounds of hierarchical merging, with domain parsing and checking, was adopted.

Evaluation of the Protein Families

The final family building procedure was benchmarked against SCOP40 (a subset of SCOP containing no sequence identities greater than 40%) version 1.63. The SCOP set includes 5226 domain sequences grouped into 1226 superfamilies and 760 folds.

As explained in methods, all pairwise detected relationships between proteins in the same superfamily were considered true positives, and all apparent relationships between proteins in different folds were considered false positives. Several other methods for detecting evolutionary relationships, BLAST, PSI-BLAST, SAM-T99 (a Hidden Markov Model method ¹¹⁵ (<http://www.soe.ucsc.edu/research/compbio/sam2src/>)), and PRC (a profile to profile method) (<http://supfam.mrc-lmb.cam.ac.uk/PRC/>), were also evaluated.

The results are shown in Figure 3.4. Overall, the new family building procedure delivers a higher fraction of true positives at low false positive rate. At the commonly adopted threshold of 1% false positives/true positives^{111; 112; 113; 114}, BLAST only detects 9% of true positives. PSI-BLAST doubles the level of detection to 18%. SAM-T99, the Hidden Markov Model Method, and PRC, a profile to profile method, both detect about 28% of true positives. Our method finds 32%, a modest but useful improvement. Note that at a higher false positive rate (above 5%, not shown in the figure), the profile-profile method performs the best. The results for BLAST, PSI-BLAST and SAM-T99 are very similar to those obtained by Park et al.¹¹⁸. Their study showed that, using the PDBD40-J dataset (similar but smaller than SCOP40), BLAST is able to detect 14% of homologous relationships and the two profile methods, PSI-BLAST and SAM-T98, can detect 27% and 29% respectively.

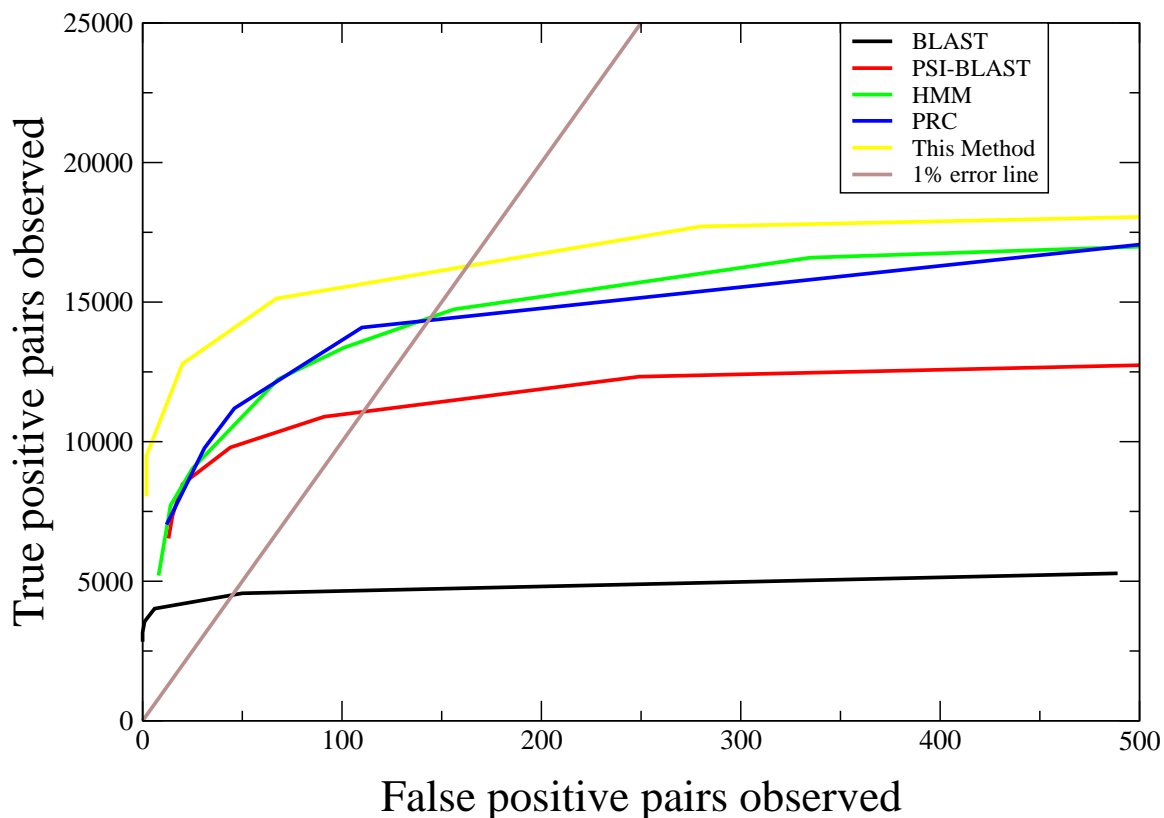


Figure 3.4. Benchmarking of the family building procedure, together with BLAST, PSI-BLAST, SAM-T99 and PRC; using SCOP40. ‘True Positive Pairs’ are the fraction of pairwise relationships within superfamilies that are detected, out of 50,374 possible. ‘False Positive Pairs’ are the fraction of apparent pair-wise relationships between folds. The more true positives detected at a given false positive rate, the better the method. At a 1% ratio of false positives versus true positives, PSI-BLAST has approximately double the sensitivity of BLAST, the simple pair-wise method. The Hidden Markov Model, SAM-T99 and the profile-profile method (PRC) improve the sensitivity to 28%. The new method achieves a modest but useful improvement to 32%. Improved sensitivity is attributed to the hierarchical linkage procedure.

3.3.2 Structural Genomics Analysis

Structure Coverage of Current Protein Families

A long term of aim of structural genomics is to obtain an experimentally determined structure for at least one protein in every family. We now ask to what extent that is already the case for the set of 67 bacterial protein families. We consider only families with three or more members, and exclude membrane protein families, since this class of structure is not yet amenable to high throughput experimental techniques. There are 4907 non-membrane protein families with three or more members. Figure 3.5 shows the fraction of families with one or more known structures (the structure coverage). About 80% of families larger than 60 have a structural representative. Coverage drops with decreasing family size, to around 5% for families with only three members. Overall, 20% of all families size three or larger have one or more representative structures. A further 3926 structures would be required to complete the coverage.

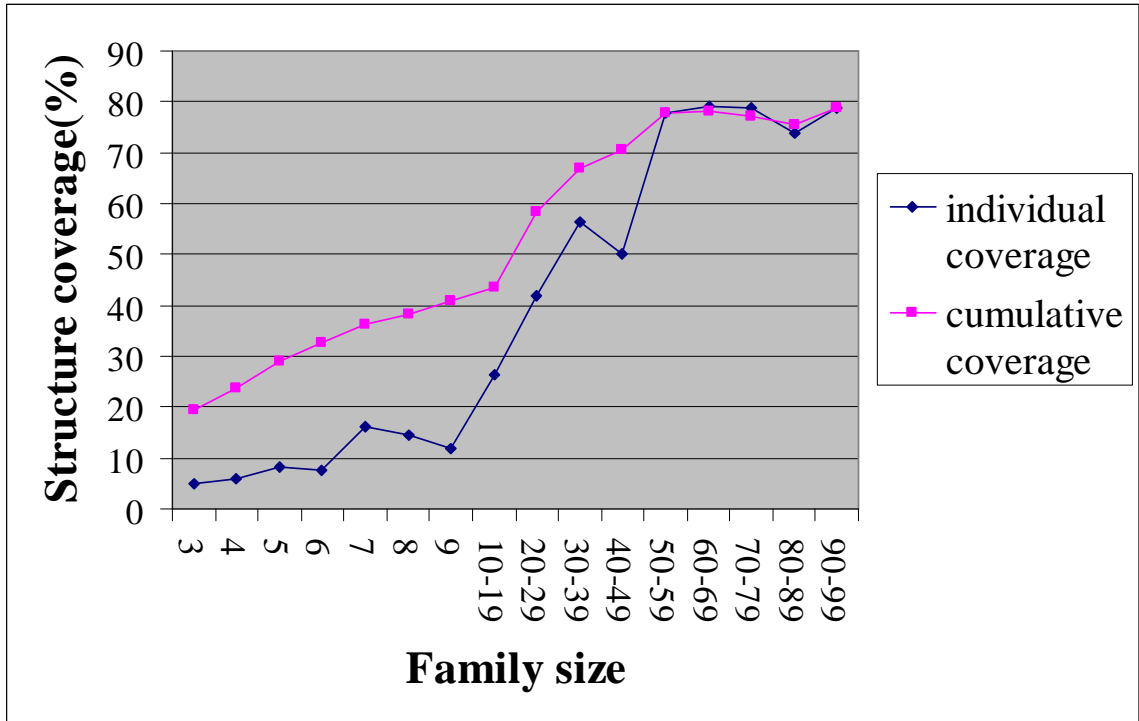


Figure 3.5. Fraction of the non-membrane protein families with three or more members for which there is at least one experimentally determined structure, as a function of family size (blue line). The purple line shows the coverage of all families that size and larger. Coverage is much larger for the larger families, approaching 80% for the biggest. The overall average coverage is 20%.

Estimation of the number of Families in a large number of Genomes

The previous section provides an estimate of the number of structures needed to complete coverage of a set of already fully sequenced genomes. Of more interest in structural genomics is the number of structures that will be needed as a function of

the number of sequenced genomes, and in particular, the limit of that quantity, I.e. the number of structures that will be needed to provide coverage of all protein families.

We have examined the increase in the number of detectable protein families as the number of fully sequenced genomes increases, using the following procedure. One of the 67 prokaryotic genomes is chosen at random, and the number of families it contains noted. A second genome is randomly selected, and the additional families present in that are added. This process is continued until all 67 have been selected. The whole procedure was repeated 100 times, and the average number of families for each number of genomes calculated.

The result is shown in Figure 3.6A. Total bar heights represent all families in the corresponding number of genomes. Subdivisions show the number of families in different size ranges, with smallest families lowest. The number of protein families is still growing rapidly up to inclusion of 67 genomes, and is far from saturation, though the rate of increase is slowing. Clearly there will eventually be many more than 30,000 detectable families. A log-log representation of these data (Figure 3.6B) is close to linear, providing a basis for extrapolation to a larger number of genomes. Figure 3.6C shows the projected number of families up to a total of 1000 genomes, using that relationship. This model predicts a total of about 250,000 families at that point, a much higher estimate than any previous ones.

The log of the number of apparent singletons also grows approximately linearly with the log of the number of genomes, with a reliability coefficient of 0.9993, and a slope of 0.704. The projected number of singletons out to 1000 genomes is shown in figure 3.6D. For a 1000 genomes, the estimate is 140,000 singletons.

The rapid growth of singletons in Figure 3.6A and the prediction made in Figure 3.6D clearly suggest their growth is also far from complete. This observation is contradictory to our earlier view that aggregation of homologs between genomes will lead to singletons' rapid disappearance⁷.

It should be born in mind that, because of the limited sensitivity of sequence methods for detecting relationships, the large number of families does not imply a similar magnitude of independent evolutionary lines.

To test the extrapolation model, we have extended the study to include 140 prokaryotic genomes (the 67 used for the extrapolation plus 73 new ones, see Methods) and built families for this set using the same procedure. The 405,709 sequences in these genomes produce 54,234 families of which 36,457 are apparent singletons. The extrapolation models predict 54,910 families and 35,807 singletons, within 1% and 2% respectively of the actual values.

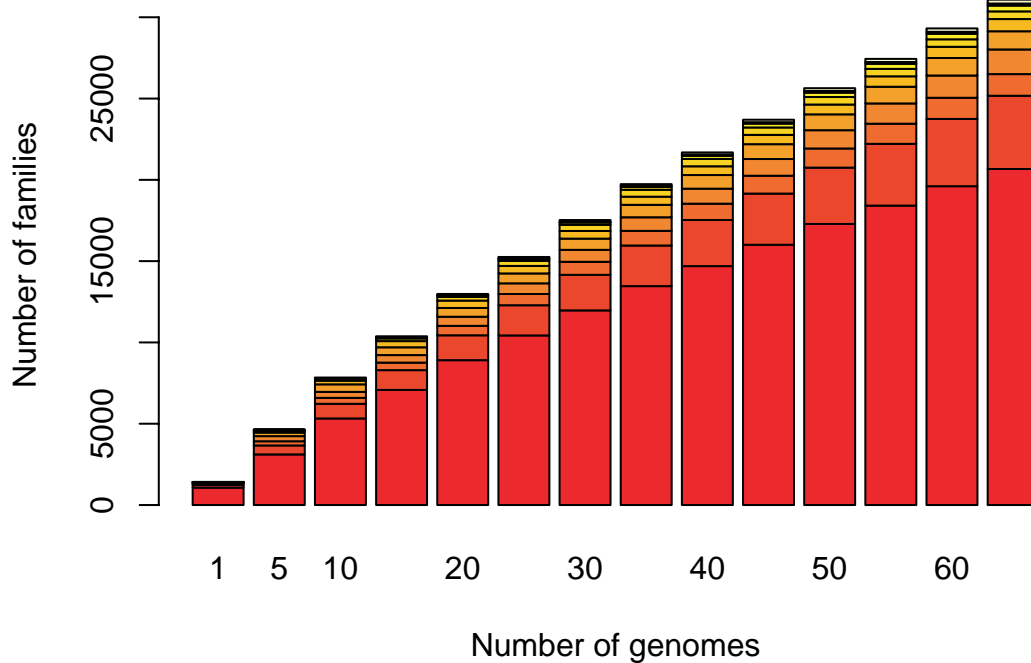


Figure 3.6A. Number of families as a function of the number of genomes. Full columns show the total families in the corresponding number of genomes, and subdivisions show the number of families in the following size ranges: 1,2,3,4-5,6-10,11-20,21-40,41-70,71-100,101-1000. Smaller families are in the lower subdivisions. The total number of families is still increasing rapidly up to 67 genomes, and is far from saturation, though there is some decrease in the rate of growth. The singleton group is the fastest grower.

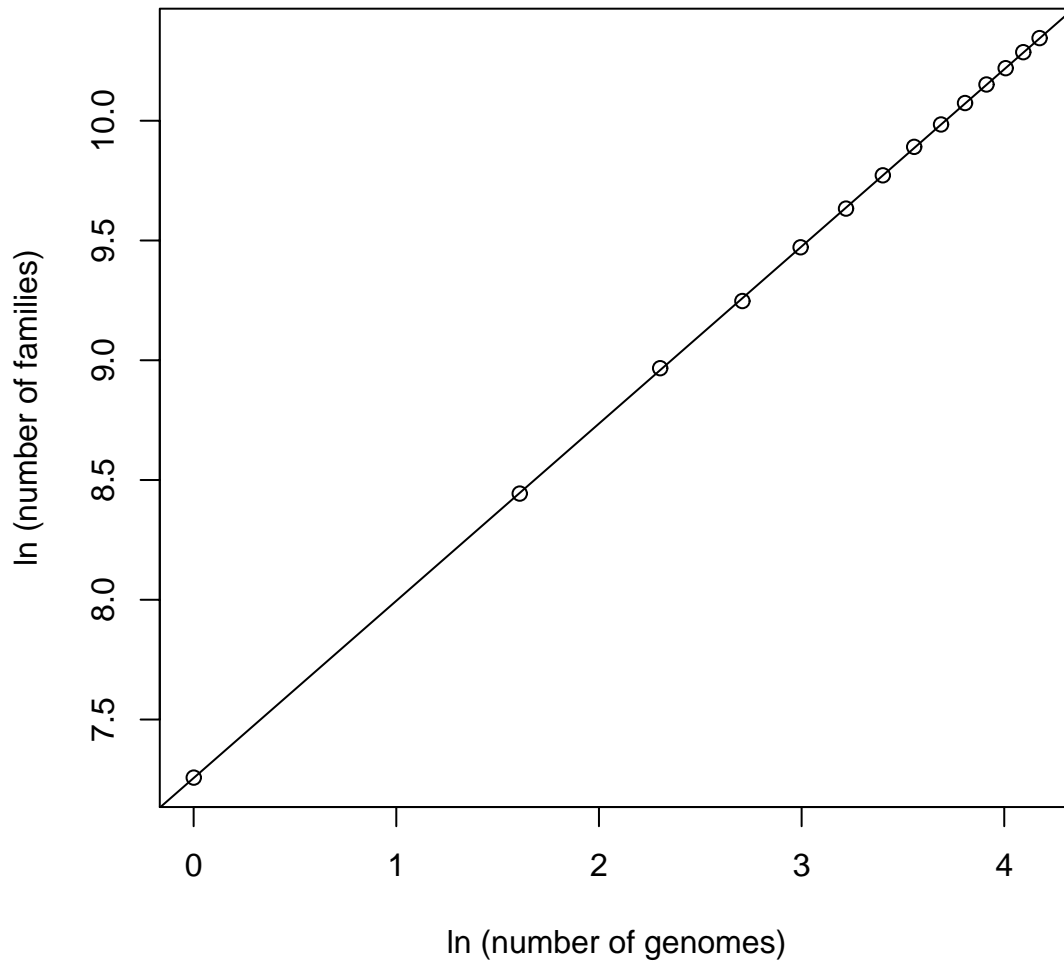


Figure 3.6B. Log-log view of the relationship between the number of families and number of genomes considered. The linear model with the slope 0.729 and the intercept 7.311 is an excellent fit to the data and the reliability coefficient is 0.9999.

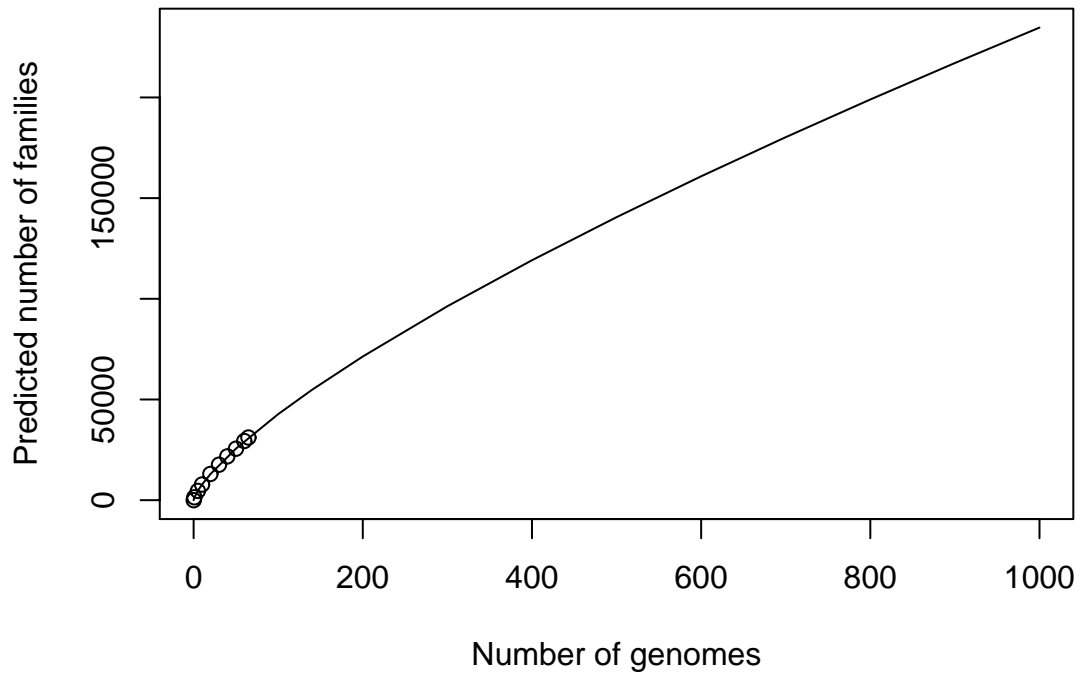


Figure 3.6C. Predicted number of families as a function of the number of fully sequenced prokaryotic genomes, based on the log linear fit in Figure 3.6B. The model predicts there will be about 250,000 families when 1000 genomes sequences are available.

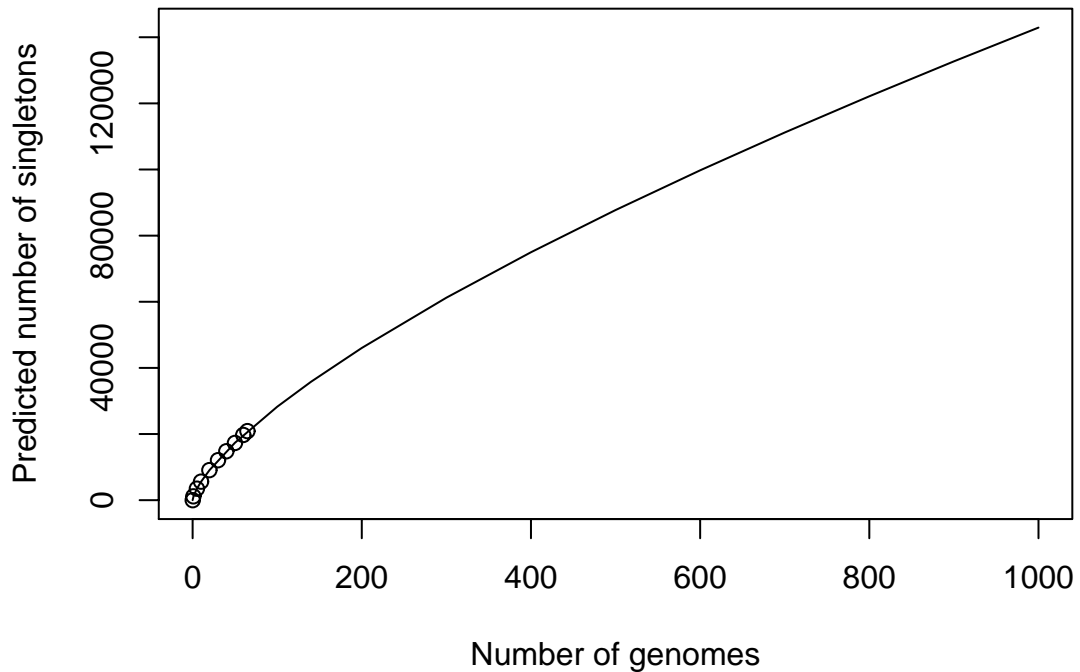


Figure 3.6D. Predicted number of apparent singletons as a function of the number of fully sequenced prokaryotic genomes. The model predicts that there will be about 140,000 when the sequences of 1000 genomes are available.

Structural Coverage for the 67 Genome Set

The previous analysis shows that it will not be possible to obtain complete structural coverage of protein family space in the near future. However, as noted earlier, a relatively small fraction of the families contain a large fraction of all the sequences. For the 67 genome analysis, 19% of the families are size 3 and larger, but contain 88% of the proteins. This suggests a strategy of obtaining representative structures for the largest families first. Figure 3.7 shows an exploration of this idea for the 67 genome set. We assume that a representative structure is first obtained for the largest

family, then the next largest, and so on. The blue curve shows the result for all non-membrane protein families with three or more members. The purple curve shows the number of structures needed, taking into account the already available structures. Because of existing high coverage, very few additional ones will be needed for large families. Altogether, about 4000 structures are required to obtain complete coverage of all families with three or more members, covering 88% of the domains in these genomes. (As discussed earlier, about 20% of these families already have representative structures). 1000 structures will complete coverage for all non-membrane families with more than 10 members, covering 80% of the domains in these genomes.

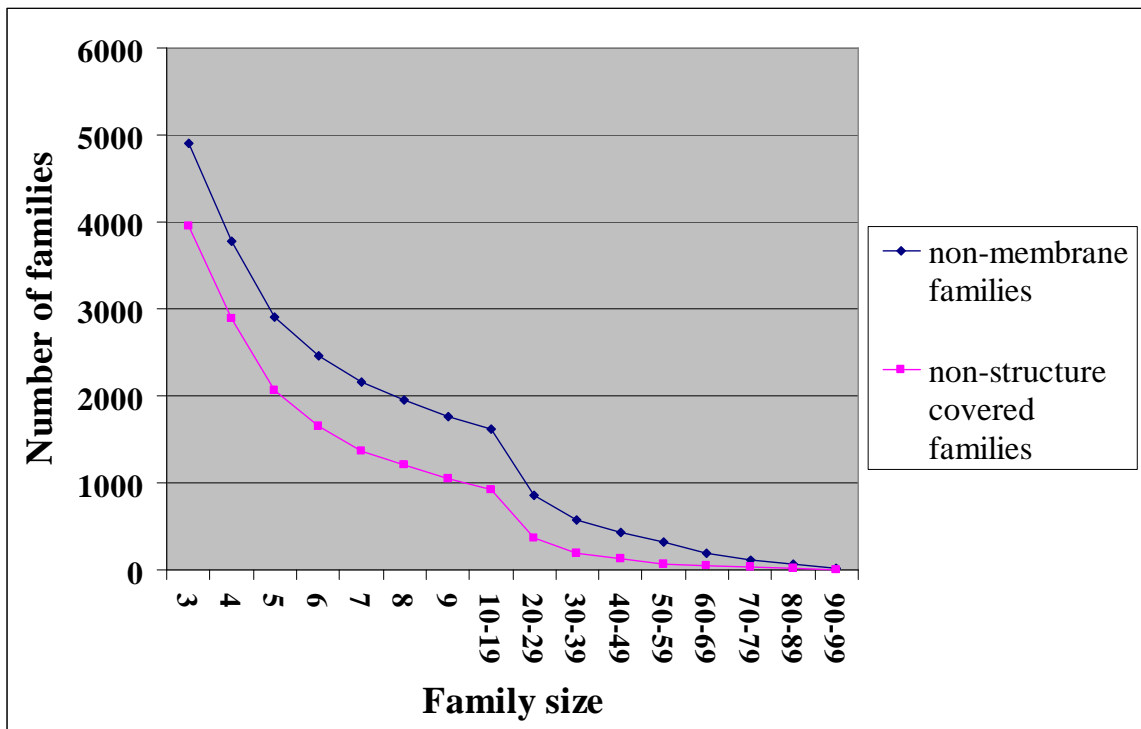


Figure 3.7. Cumulative number of experimental structures needed to obtain complete coverage of families size 3 and larger, starting with large families (right side of the

plot). The blue curve is for all non-membrane protein families, and the purple one for those families with no current structural coverage. Very few additional structures are needed to complete coverage of large families: 1000 optimally selected ones would complete coverage of all non-membrane families larger than 10, including 80% of all the domains. About 4000 would be needed to provide one structure per family size three or larger, and would cover 88% of all the domains. (These numbers are for the set of 67 genomes analyzed in this work).

Achievable Structural Coverage for 1000 Genomes

We now examine how many structures will be needed to achieve a given level of protein coverage, as the number of fully sequenced genomes grows. For that purpose, a similar extrapolation procedure to that described earlier was used. A genome was picked at random from the set of 67. The number of families was then calculated for that genome alone. The number of structures needed to obtain coverage of various fractions of all the proteins in that genome was calculated, assuming structures for the largest families are obtained first. Another genome was then randomly selected, and the number of structures needed to obtain various fractions of domain coverage for the two genomes was calculated, and so on, up to 65 genomes. The simulation was repeated 100 times, and the results averaged, to remove bias in genome order.

Figure 3.8A shows the results. Here, 100% coverage implies models for all domains in all families, 90% that 90% of the domains will have models, and so forth. The general trend is that the lower the domain coverage required, the slower the growth of the number of structures needed, as a function of the number of genomes. The growth rates for 80 and 90% coverage are already decreasing when 65 genomes are considered, and growth has almost ceased for 50% and 60% coverage. Figure 3.8B shows the estimates for up to 1000 genomes, based on log linear models. At that stage, less than half of the number of structures are needed for 90% coverage as for 100%, and the growth rate for 70% or lower coverage is slow. Figure 3.8C shows an expansion of the region below 80% coverage. Representative structures for about 8000 families will provide 70% coverage of all the domains in a 1000 genomes. This is a reasonable expectation for the next decade, given the rate of accumulation of new experimental structures.

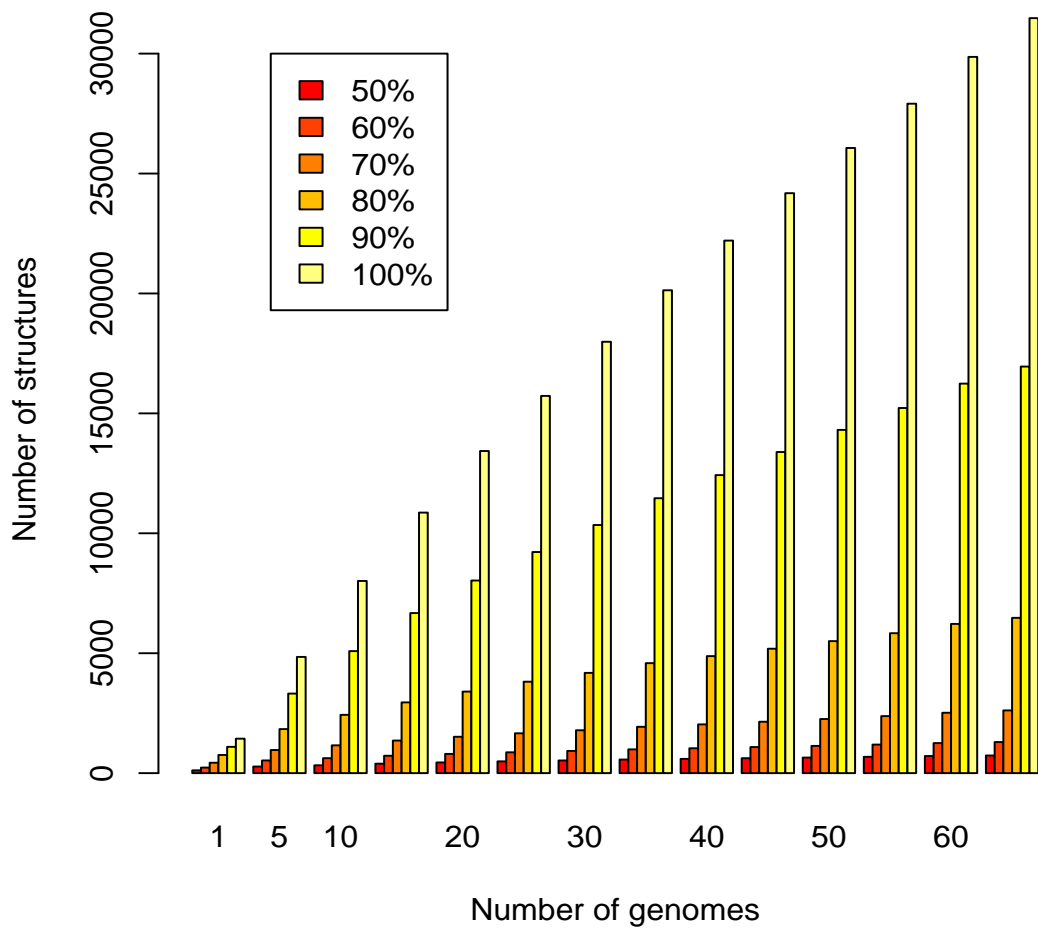


Figure 3.8A. Number of families with representative structures needed to provide structural coverage for different fractions of protein domains, as a function of the number of fully sequenced genomes. The lower the domain coverage required, the slower the growth in the number of families.

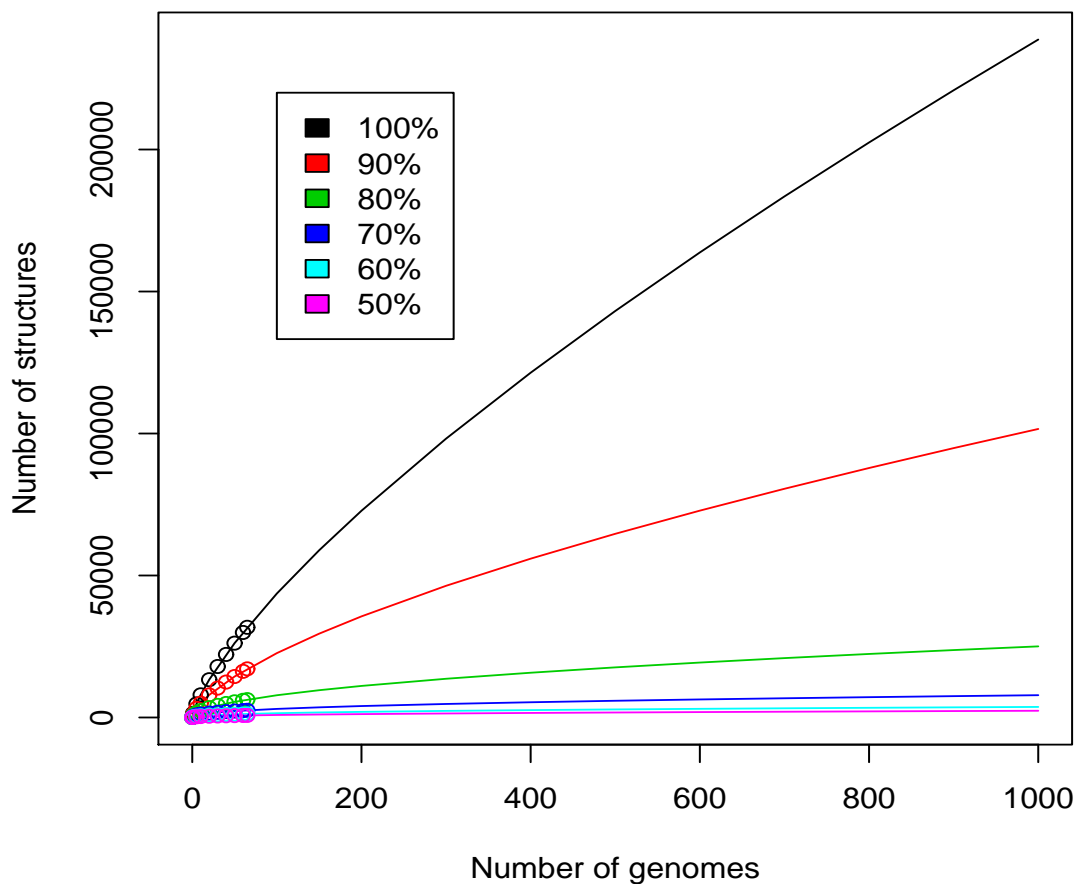


Figure 3.8B. Projection of the number of families with representative structures needed to obtain structural coverage of different fractions of protein domains, up to 1000 genomes. 250,000 structures would be required to obtain 100% coverage of these families, but 90% coverage would be obtained for less than half of that number.

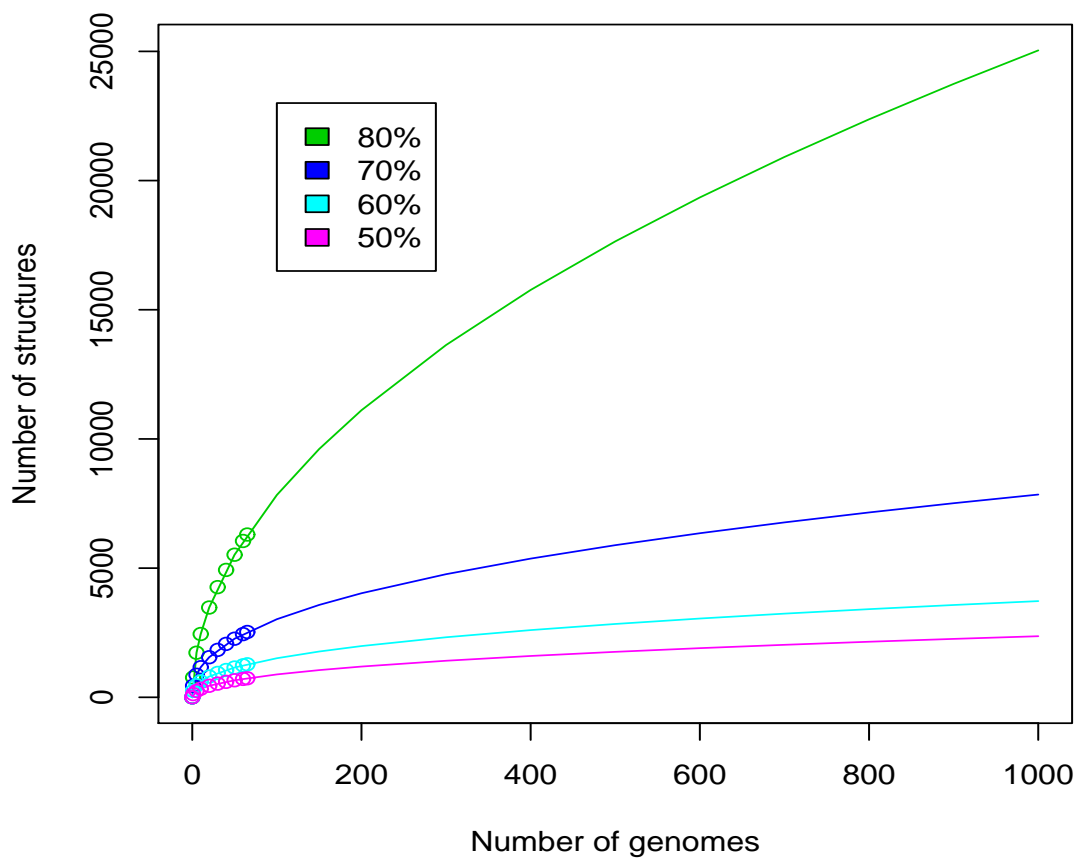


Figure 3.8C. Expansion of Figure 3.8B for coverage between 50% and 80%. For 1000 genomes, approximately 8,000 structures are needed to provide 70% domain coverage, achievable in the next decade, considering the rate of accumulation of solved structures.

3.4 Conclusion and Discussion

A principal goal of structure genomics is to obtain structures for a large fraction of naturally occurring proteins. This goal can be achieved by experimentally determining at least one structure for each protein family and building structure models for all other proteins, using comparative modeling methods¹¹⁹. The minimum number of experimental structures required for complete structural coverage of protein space is then equal to the number of apparent protein families. In a previous study⁷, we estimated this number by analyzing PfamA families⁴, and making a very simple extrapolation of likely future growth in the number of families.

In the present study, we have based the analysis on all families in a set of fully sequenced prokaryotic genomes, rather than the contents of PfamA. A major difference is the inclusion of all proteins, not just those in the larger families typically collected in PfamA. With this more realistic view of the protein universe, we find there are a very large number of such small families: for the set of 67 genomes analyzed, there are 25,802 families with only one or two members, out of a total of 31,874. Overall, there is an approximately power law relationship between the number of families and family size.

Use of complete genome sequence sets has also allowed us to use a more realistic extrapolation method, in order to estimate the future growth in the number of

families, as the number of fully sequenced genomes grows. We find that when 1000 genomes are available, there will be about 250,000 detectable protein families. Further, the number of families will still be growing at that point.

The large number of families makes it clear that complete structural coverage of protein space will not be possible in the near future. Nevertheless, it will be possible to obtain structural models for a high fraction of proteins. This is because most proteins belong to large families – for the 67 sequenced genomes, 88% of the proteins fall into just 6,072 families. Further, the extrapolation model shows that this trend will continue, so that, considering all sequences, 80% structural coverage of the proteins in 1000 genomes can be obtained with 25,000 structures, and 70% coverage with 8,000. The primary conclusion from this work is that a strategy of obtaining structural representatives for the largest families first will lead to a high fraction of structural coverage of protein space within the next decade. This strategy will also lead to early structural coverage of the families that perform more universal biological functions, and will provide the most leverage of experimental effort, by creating models for the largest number of proteins from each experimental structure. We envisage that when structures for proteins in small families are needed, they will typically be obtained one at a time, using conventional structural biology, rather than high throughput methods.

The number of apparent protein families depends on the effectiveness of each of the steps in building them. There are three key steps in our procedure. The first step uses

PSI-BLAST to search for relatives of each protein. Other methods, in particular well tuned Hidden Markov Models ¹¹⁵ and profile-profile methods ^{95;97 96} are more sensitive for this purpose ^{118;95}. It turns out that the later merging step compensates for PSI-BLAST's relative insensitivity.

The second step of family building is parsing of proteins into domains. We have used a sequence profile based approach, relying on the fact that the most insertions and deletions occur between domains ^{88 92}. We apply the procedure very conservatively to minimize splitting within domains. As a consequence, this step has many false negatives – it does not split at many domain boundaries that are obvious at the structure level. The accuracy of the method is similar to that of CHOP ¹²⁰, although those authors chose a different compromise between false negatives and false positives in their analysis. Two additional procedures might further improve our method: Mapping known structural domains and PfamA (hand curated) domains onto the proteins. We have not done that because the majority of these families have not yet been studied structurally, and many are not yet in PfamA, so that use of these signals may distort the choice of parameters for family building.

The third step in family building is merging lists of related domains and filtering out redundant entries, to create domain families. As noted earlier, over-merging is a well known problem in protein family building – a small number of incorrect entries in the initial lists of relatives can easily lead to substantial over-merging. To avoid this, we

use a procedure that requires an increasing number of common entries as a function of alignment size.

The rules for merging and other steps were tuned by reconstructing a set of PfamA domains from the corresponding full length sequences, and comparing the generated families with the PfamA ones. The final procedure was benchmarked by comparing pair-wise relationships within a set of generated families with those in a set of SCOP superfamilies. While these testing methods are very useful, they are not ideal. PfamA is a sequence based family set, and so omits a large number of evolutionary relationships (placing related proteins in different families). A more sensitive method may therefore appear to have an excessively large number of false positives, and consequently may be detuned to reduce these. PfamA also focuses on larger families, whereas the genome data is dominated by smaller families. As a result, a better method for PfamA may not necessarily be optimum on genome data, and performance may be different from that suggested by quality measures on PfamA. Similarly, SCOP contains only proteins with known structure, and these may be unrepresentative of proteins in genomes as a whole, for example not included proteins with significant inherent disorder, and under-representing proteins that form part of complexes. Nevertheless, PfamA and SCOP are probably the best training and test sets available. As in most of computational biology, the lack of a gold standard for methods development and evaluation is an inherent limitation.

According to our and other benchmarking, at a 1% ratio of false positives versus true positives, only about 30% of the pair-wise evolutionary relationships implied by structure can be recovered with present sequence comparison methods. A consequence is that families built with those methods do not approximate independent evolutionary lines. As more structures are available, the number of families will decrease very substantially, because of merging on the basis of structural similarity, rather than sequence. For the purposes of structural genomics, a single representative structure for very large families containing very remote relatives is not particularly desirable. As the remoteness of the relationship between proteins increases, the quality of a model built on the basis of a relative with an experimental structure decreases. In particular, a substantial fraction of residues (up to 50%) will have no equivalent in the modeling template ¹²¹. Thus, although families generated from sequence relationships are suboptimal from an evolutionary standpoint, they are very suitable for structural genomics.

Contrary to our earlier expectation ⁷, the number of apparent singletons and other small families will continue to increase. Siew and Fischer also found the number of singletons is steadily growing, though the percentage of singletons as a fraction of all sequences is decreasing ⁵. Because of the limited sensitivity of sequence methods, it is not possible to judge the biological significance of this at present. Many singletons may in fact have unrelated folds, as one estimate of the total number of folds suggests ¹²². Or, most may turn out to be members of larger superfamilies, too remotely related

for sequence methods to detect. A larger set of experimental structures of small families will settle this issue.

This work assessed how many experimental structures will be needed to provide models of a given fraction of all naturally occurring domains, based on one representative structure per family. Under this strategy, the majority of structures will be domain models, based on a single experimental within a protein family. While such a structure set will revolutionize our view of proteins in many ways, it is only the first step in providing complete structural information for natural proteins. Many proteins are multi-domain, particularly in higher eukaryotes^{123; 124}, and the function of a domain assembly is not always a simple combination of that in the constituent domains¹²⁴. Generating reliable multi-domain structures will sometimes involve docking of domain models, requiring improvements in computational methods, or further experimental structures. Second, the relationships within families on which models will be based are often fairly distant, with sequence identities well below 30%. Models based on such sequence relationships contain substantial errors, primarily arising from mistakes in aligning the sequence of interest with those of available templates, and because significant parts of the structures will differ from that of the templates⁸². Nevertheless, these low accuracy models will be adequate for establishing membership of a superfamily, and thus useful for a variety of purposes, including providing approximate molecular function information, guiding site directed mutagenesis experiments, and choosing likely antigenic peptides. Other uses, such as identification of ligand specificity¹²⁵ and interpretation of the effect of

disease related mutations ¹²⁶, require higher accuracy, only possible by modeling against a template with 30% or higher sequence identity. Comprehensive structural information at that level will require many more structures.

Chapter 4: Lateral Gene Transfer between Prokaryotic Genomes

4.1 Introduction

Lateral gene transfer, also called horizontal gene transfer, is the process of transfer of genes between different species. Its significance was not appreciated until the 1950s, when resistance to penicillin class antibiotics spread rapidly through many pathogens as a result of plasmid transfer⁹. For many years thereafter it was still commonly believed that lateral gene transfer was rare, and did not play a significant role in evolution. As sequenced-based genomics has developed, it has become more and more obvious that the process is very common and plays an important role in evolution. It is now clear that in prokaryotes, it acts as a significant force in the diversification of species¹²⁷.

Successful lateral gene transfer requires three steps. First, donor DNA must be delivered to a recipient cell. There are several possible mechanisms:

1. Transformation. The uptake of naked DNA from the environment.^{128; 129; 130}

Transformation is likely very inefficient comparing with other processes.

2. Phage transduction. Phage replication in a donor cell results in the incorporation of some donor genome fragments. Subsequently, the phage is transmitted and absorbed by the recipient cell. This can only happen between two species both within the infection spectrum of the phage. The size of fragments transferred is limited by the size of the phage capsid, but can be up to 100kb^{131; 132; 133; 134}. As with transformation, phage transduction does not require physical contact between donor and recipient cells.

3. Conjugation. Genetic material is transferred between donor and recipient when two cells are in physical contact. This can happen between distantly related species¹³⁴.

Second, the acquired genetic information must be incorporated into the recipient cell's genome. Mechanisms involved include:

1. Transposon mediated transfer. A transposon contains segments of nucleotide sequence flanking the two ends of the transferred material and can help move it to different locations in a genome or different genomes. Transposons are generally moved in a "cut and paste" way: the transposon is cut out of its original location and inserted into a new location. This process requires an enzyme, a transposase, encoded within some transposons.^{135; 136; 137}

2. Phage transduction. A phage may also assist the integration of foreign DNA segments into a chromosome, using an integrase.

Third, the transferred sequence must be expressed in the recipient cell in a manner that potentially benefits the recipient organism. While the first two steps are largely unrelated to the function and properties of the transferred genes, the third step is subject to natural selection.

Many previous analyses have identified individual examples of lateral gene transfer (LGT) ^{138; 139}. Knowledge of the complete genome sequences of a large number of organisms provides new opportunities for a more global view of the extent and nature of the process. It has been estimated that between 8% and 18% of the *E.coli* genome was acquired by lateral gene transfer ^{127; 140; 141}. In other genomes, the estimated extent of transfer varies over a wide range, from almost none in small genomes such as *Mycoplasma genitalium*, *Rickettsia prowazekii* and *Borrelia burgdorferi*, to about 24% in *Thermotoga maritima* ^{134; 142; 143}. Studies have shown that lateral gene transfer events can happen across large phylogenetic distances, for example, isoleucyl-tRNA synthetases, whose acquisition from eukaryotes by several bacteria is linked to antibiotic resistance ¹⁴⁴. Transfers from eukaryotic to prokaryotic organism happen rarely. Transfers from prokaryotes to eukaryotes are even less likely, presumably because only transfers into eukaryotic germlines are potentially viable ^{126; 145}.

Transferred genetic material may help a host acquire new function capabilities, and thereby promote fitness and adaptation ¹⁴⁶. Well-known cases include antibiotic resistance ⁹, pathogenicity islands ^{133; 147}, novel metabolic capabilities ^{15; 134; 148} and non-

orthologous gene replacement¹⁴⁸. Furthermore, accumulated differences introduced by lateral gene transfer can prompt species divergence and new species formation^{134; 140}.

Two methods for identification of lateral gene transfer have been developed. The first makes use of the fact that gene compositions such as GC content and codon usage bias vary significantly between species¹⁴⁹. Thus, it is in principle possible to detect genes that have recently been transferred to an organism with sufficiently different GC and codon properties^{140; 150; 151}. Statistical methods have been employed to quantitatively assess the composition of individual genes against the genome signature. Garcia-Vallve et al¹⁵¹ considered genes as foreign when their whole GC content deviated by $> 1.5\sigma$ from the genome mean value or when the GC content in the first and third codon positions have the same deviation direction from the genome mean and at least one of them is $> 1.5\sigma$. Lawrence and Ochman¹²⁷ and Sharp et al.¹⁵² made use of abnormal patterns of codon usage to identify transfer events. A limitation of these methods is that laterally transferred genes cannot be identified if the donor and recipient organisms have similar GC and codon use profiles. Although the methods are simple and straightforward, Koski et al. found that composition measures may not be a reliable indicator of horizontal transmission¹⁵³. These authors observed that a number of *E.coli* native genes with intrinsic atypical compositions were incorrectly classified as lateral transferred. Lawrence and Ochman also pointed out the difficulty of identifying ancient transfers by composition methods, since the nucleotide composition and codon use of transferred genes drifts towards that of the new host, so called “amelioration”^{134; 141}.

The second existing method for identifying lateral gene transfer is analysis of incongruence between the phylogenetic trees of genes.^{154; 155} Generally, a lateral gene transfer event within a protein family can be inferred when two of the sequences have anomalously high sequence identity, resulting in a family phylogenetic tree that differs from the species based one. Comparison of tree topologies is a manual process, prohibiting large scale lateral transfer screens by this method, and the lack of a quantitative measure further complicates the approach.

We have developed two new methods for identifying lateral transfer events. The first, the High Apparent Gene Loss method (HAGL), makes use of the fact that a lateral transfer event will introduce a number of apparent gene losses in the conventional phylogenetic tree of a protein family. To appreciate this, consider a case of transfer of a proteobacterial gene into a single genome in the archaeal kingdom. A conventional evolutionary inheritance interpretation will imply that the ancestral gene has been lost in all other archaeal genomes. The higher the number of implied losses relative to the protein family size, the more likely that a lateral gene transfer event has occurred. The minimum number of losses needed to explain the observed distribution of family members over genomes is derived using the Dollo Maximum Parsimony algorithm¹⁵⁶(Farris, JS. Phylogenetic analysis under Dollo's law. 1977. *Syst Zool*, 26, 77-88)(Le Quesne, WJ. 1974. The uniquely evolved character and its cladistic application. *Syst. Zool.* 23 513-517). This method is particularly effective at identifying transfer within small protein families.

The second new method, termed the Evolutionary Rate Anomaly method (ERA), identifies LGT events by finding those proteins which exhibit an anomalous rate of sequence change. Sequence differences between pairs of proteins in different species are used to derive an estimated number of accepted substitutions per amino acid position since species divergence. These accepted substitution levels are converted to relative rates of substitution by dividing by the corresponding mean substitution level in a set of highly conserved protein families. Gene transfer between species results in a lower apparent rate of accepted substitution, providing a means of identifying LGT events. Two factors complicate interpretation: Accepted substitutions between proteins that are in paralogous subfamilies are typically larger than expected, and the rate of evolution within particular protein families is not always constant. We largely eliminate the first factor by only performing the analysis on apparently orthologous subfamilies. We identify uneven evolutionary rate cases by examining the consistency of pairwise substitution levels using a modified version of a robust linear regression procedure: Least Median of Squares¹⁵⁷.

The new methods have been applied to analysis of lateral gene transfer using 66 fully sequenced prokaryotic genomes. Both methods require that proteins be grouped into families. For this purpose, we make use of family building procedures described elsewhere (Yan and Moulton, Protein Family Clustering for Structural Genomics, submitted) to establish a set of protein families. The High Apparent Gene Loss method works best for small families where there has been transfer over large phylogenetic distances. The Evolutionary Rate Anomaly method works best for larger families with a

steady rate of evolution. In the absence of a reliable set of known LGT cases, the methods have been calibrated and evaluated against each other.

Results from both methods confirm that lateral gene transfer events are widespread in prokaryotic genomes. Over-all, 18% of the genes analyzed are classified as transferred. Together, the two methods only identify a subset of all LGT events, suggesting that the scale of lateral transfer events is even larger. Analysis of the results in terms of families and genomes indicate that transfer has occurred unevenly. Many large protein families appear to have no lateral transfer events, whereas transfer was found in many small protein families. For genomes, values vary greatly, from 3% of analyzed proteins in *Mycoplasma genitalium* and *Buchnera sp. APS*, to 33% in *Nostoc sp. PCC7120*, a Cyanobacterium and *Halobacterium sp. NRC-1*, an archaeon.

4.2 Methods

Protein Sequences in microbial genomes:

Complete sets of protein sequences for all genomes were retrieved from Genbank (<http://www.ncbi.nlm.nih.gov/Genomes/index.html>). All downloaded and generated information were stored in a MySQL relational database running on a Linux server.

Genome	Genome abbreviation	Number of proteins
<i>Aeropyrum pernix</i>	Aero	2694
<i>Agrobacterium tumefaciens str. C58 (Dupont)</i>	Atum_D	5402
<i>Aquifex aeolicus</i>	Aquae	1553
<i>Archaeoglobus fulgidus</i>	Aful	2407
<i>Bacillus halodurans</i>	Bhal	4066
<i>Bacillus subtilis</i>	Bsub	4100
<i>Borrelia burgdorferi</i>	Bbur	1637
<i>Brucella melitensis</i>	Bmel	3198
<i>Buchnera sp. APS</i>	Buch	574
<i>Campylobacter jejuni</i>	Cjej	1629
<i>Caulobacter crescentus</i>	Ccre	3737
<i>Chlamydia muridarum</i>	ctraM	916
<i>Chlamydophila pneumoniae AR39</i>	cpneuA	1110
<i>Chlamydophila pneumoniae CWL029</i>	cpneuC	1052
<i>Chlamydophila pneumoniae J138</i>	cpneuJ	1069
<i>Clostridium acetobutylicum</i>	Cace	3672
<i>Clostridium perfringens</i>	Cper	2723
<i>Corynebacterium glutamicum</i>	Cglu	3040
<i>Deinococcus radiodurans</i>	Dra	3102
<i>Escherichia coli</i>	Ecoli	4289
<i>Escherichia coli O157:H7</i>	ecoliO157	5361

<i>Haemophilus influenzae Rd</i>	Hinf	1709
<i>Halobacterium sp. NRC-1</i>	Hbsp	2605
<i>Helicobacter pylori 26695</i>	Hpyl	1566
<i>Helicobacter pylori J99</i>	Hpyl99	1490
<i>Lactococcus lactis subsp. lactis</i>	Llact	2266
<i>Listeria innocua</i>	Linn	3043
<i>Listeria monocytogenes EGD-e</i>	Lmon	2846
<i>Mesorhizobium loti</i>	Mlot	7275
<i>Methanobacterium thermoautotrophicum</i>	Mthe	1869
<i>Methanococcus jannaschii</i>	Mjan	1770
<i>Mycobacterium leprae</i>	Mlep	1605
<i>Mycobacterium tuberculosis</i>	Mtub	3869
<i>Mycobacterium tuberculosis CDC1551</i>	Mtub_cdc	4187
<i>Mycoplasma genitalium</i>	Mgen	480
<i>Mycoplasma pneumoniae</i>	Mpneu	688
<i>Mycoplasma pulmonis</i>	Mpul	782
<i>Neisseria meningitidis</i>	Nmen	2025
<i>Neisseria meningitidis Z2491</i>	nmenA	2032
<i>Nostoc sp. PCC 7120</i>	Nost	6129
<i>Pasteurella multocida</i>	Pmul	2014
<i>Pseudomonas aeruginosa</i>	Paer	5565
<i>Pyrobaculum aerophilum</i>	Paero	2605
<i>Pyrococcus abyssi</i>	Pabyssi	1765

<i>Pyrococcus horikoshii</i>	Pyro	2064
<i>Ralstonia solanacearum</i>	rsol	5116
<i>Rickettsia conorii</i>	Rcon	1374
<i>Rickettsia prowazekii</i>	Rpxx	834
<i>Salmonella enterica subsp. enterica serovar Typhi</i>	Sent	4749
<i>Salmonella typhimurium LT2</i>	Styp	4553
<i>Sinorhizobium meliloti</i>	Smel	6205
<i>Staphylococcus aureus subsp. aureus Mu50</i>	Saur_mu50	2748
<i>Staphylococcus aureus subsp. aureus N315</i>	Saur_n315	2624
<i>Streptococcus pneumoniae</i>	Spneu	2094
<i>Streptococcus pyogenes</i>	Spyo	1696
<i>Sulfolobus solfataricus</i>	Ssol	2977
<i>Sulfolobus tokodaii</i>	Stok	2826
<i>Synechocystis PCC6803</i>	Synecho	3169
<i>Thermoplasma acidophilum</i>	Tacid	1478
<i>Thermoplasma volcanium</i>	Tvol	1526
<i>Thermotoga maritima</i>	Tmar	1846
<i>Treponema pallidum</i>	Tpal	1031
<i>Ureaplasma urealyticum</i>	Uure	611
<i>Vibrio cholerae</i>	Vcho	3828
<i>Xylella fastidiosa</i>	Xfas	2831
<i>Yersinia pestis</i>	y pes	4039

Table 4.1. The 66 fully sequenced microbial genomes used in the Lateral Gene Transfer analysis, and the number of proteins in each genome. 12 are archaeal, and 55 are bacterial. In total there are 178,310 protein sequences.

Generation of a Domain-based Protein Family Set

Following the procedure described in [Yan and Moulton, Protein Family Clustering for Structural Genomics, submitted], the 178,310 sequences proteins were parsed to 249,574 domains, and then clustered into 31,874 homologous sequence families. Small families predominate. In particular, there are 20,992 singletons (families with only one member), about 2/3 of the total. The 6072 protein families containing three or more members are used in this analysis.

Extraction of Orthologous subfamilies

In many cases, a single family represents more than one function, as a result of gene duplication and specialization (for example, malate and lactate dehydrogenases are grouped in a single family). These paralogous subfamilies may evolve at different rates and there are often rapid sequence adaptations associated with function change, so it is desirable to divide families into orthologous subfamilies.

Orthologous subfamilies were extracted as follows:

1. A sequence identity matrix is generated for each family. Multiple sequence alignments are generated using CLUSTALW¹⁵⁸. Pair-wise sequence identities are calculated from the alignments, providing the matrix elements.

2. A kernel protein is chosen. For each sequence in the family, sequence identities scores to all other sequences were summed. The kernel protein is the one with the highest score, at the arithmetic center of the family. The kernel protein provides the starting sequence for building an orthologous subfamily.

3. Additional proteins are added iteratively. The closest protein sequence (highest sequence identity to the kernel protein) from another genome is first added to the orthologous group. Then the average sequence identity of each protein in the remaining genomes to those already in the subfamily is calculated, and the closest one added. This step is repeated until a representative sequence from each genome is included or the average sequence identity between any remaining candidate protein and the orthologous group is less than 15%. (The 15% threshold reduces the possibility of including poor alignments or incorrect family members).

4. Steps 1 through 3 may be repeated for the remaining proteins, until none are left, to generate further orthologous sub-families. The first sub-families extracted are the most reliable. So lateral gene transfer analysis has been performed only on these. The 6072 initial families with three or more members produced 4856 orthologous families size three or larger.

Phylogenetic Tree Construction

A universal phylogenetic species tree provides an approximation to the true evolutionary relationships among species, and is a useful reference in our analysis of LGT events. In the HAGL method, the topology of the species tree is used to estimate the minimum number of gene loss events required to explain the phyletic pattern (presence and absence in phylogenetic lineages)¹⁵⁹ of a protein family.

A number of inter-species metrics have been adopted in constructing species trees, including the widely used 16S ribosomal RNA sequence identity¹⁶⁰, the Common Gene Fraction and Common Neighbor Fraction (Yan and Mout, 'Operon Predictions in Microbial Genomes', submitted)⁵⁷ and the average sequence identities over a set of conserved orthologous protein families^{161; 162}. Two reference trees were considered in this work: the 16S rRNA tree and a conserved protein families tree. 16S ribosomal RNA sequence data were downloaded from the Ribosomal Database Project (<http://rdp.cme.msu.edu/>) and a 16S rRNA distance matrix was built using the DNADIST program in the PHYLIP package, release 3.6 (Felsenstein 1989)³¹ For the conserved protein families tree, a set of fourteen conserved orthologous protein families (listed in Table 4.2), all with members in each of the 66 genomes, were chosen. Interspecies protein distance matrices were calculated for each family, using the PROTDIST program in PHYLIP package (Felsenstein 1989)³¹. The Jones-Taylor-Thornton matrix amino acid substitution model¹⁶³ was used to obtain the estimated average substitution per amino

acid from sequence identities. Matrices for the 14 families were averaged to obtain the final interspecies matrix.

Protein	Amino Acids
Ribosomal protein L14	125
Ribosomal protein L13	149
Ribosomal protein S17	93
Ribosomal protein L2	271
Ribosomal protein S2	254
Ribosomal protein L5	181
DNA-directed RNA polymerase, alpha subunit	333
Ribosomal protein L10	200
Ribosomal protein S13	128
Ribosomal protein S5	185
Ribosomal protein L15	147
Preprotein translocase secY subunit	447
Ribosomal protein S3	234
Ribosomal protein S11	130

Table 4.2. Fourteen well conserved orthologous protein families used to generate the average interspecies distance matrix. Most are ribosomal proteins. The average number of amino acids in each protein family is also shown.

Several tree building methods are available. Brown et al ¹⁶¹ compared trees constructed using Maximum Likelihood, Neighbor-Joining, Maximum Parsimony and a Minimum Evolution method for 16S rRNA and a set of conserved orthologous protein families. Their study found these trees to have highly congruent topologies. We built protein and 16S rRNA species trees from the distance matrices described above, using the Neighbor Joining method ³¹ as implemented in the NEIGHBOR program (Felsenstein 1989) ³¹. As shown later, the trees are very similar topologically.

The High Apparent Gene Loss method (HAGL)

For protein families, the phyletic pattern (presence or absence of family members in the organisms considered) is used to deduce the minimum number of gene loss events that occurred during evolution, assuming only a single ancient gain event, and a classical pattern of inheritance for all family members. Figure 4.1 shows a schematic example, for a protein family with only two members. One of the members belongs to species A, the other belongs to species F and all other species have no members of this family. Presence of a family member at a node in the tree is represented by a '1', and absence by a '0'. Given this phyletic pattern, and the reference phylogenetic tree, the Dollo Maximum Parsimony algorithm will find the minimum number of losses consistent with these data, assuming a single ancestral gain event. The method was first suggested by Le Quesne (1974) and named after Louis Dollo, since he was one of the first to assert that in evolution it is harder to gain a complex feature than to lose it. In this case, Maximum

Parsimony requires three loss events (state 1--> 0, shown by 'X's) to explain the phyletic pattern, assuming only one ancestral gain event (state 0-->1) some time prior to the closest common ancestor ('CA' in the figure).

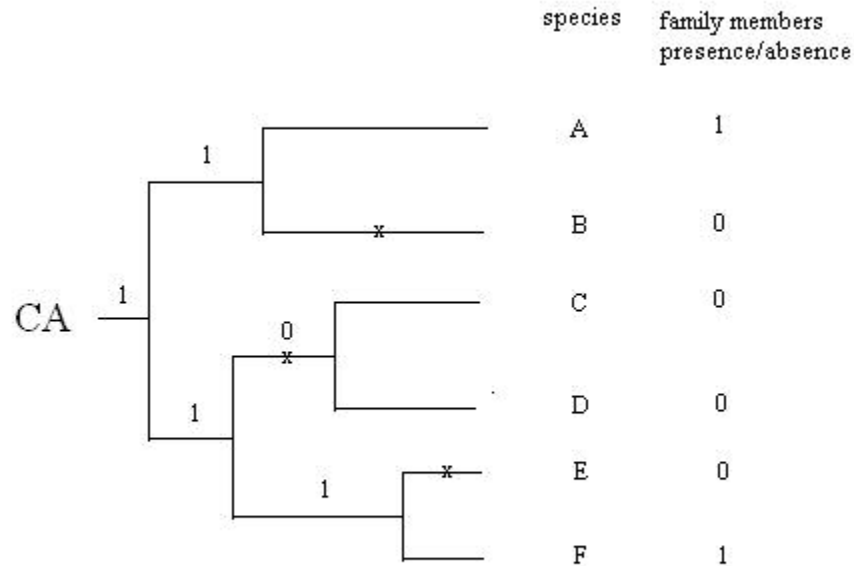


Figure 4.1: Phyletic pattern of a hypothetical protein family in a species tree. The family has two members, belonging to species A and F (indicated by '1' states). Assuming a single ancient gain event, the minimum number losses necessary to explain the pattern in terms of a standard evolutionary process is three, in the branches marked with 'x'. 1's and '0's on the branches show the inferred presence or absence of ancestors. 'CA' represents the closest common ancestor of this family.

An alternative explanation for the observed phyletic pattern is that there have been one or more lateral gene transfer events. Most simply in this case, a gain event in either the A or F genomes was followed by a transfer to the other one. Less simply, there may have been two independent transfer events from a third, unsequenced genome.

The more losses required to explain a phyletic pattern in terms of evolutionary descent, the more likely the correct explanation involves gene transfer. The simplest possible model assumes a constant probability of a loss event per unit branch length in the tree. Then, the more ancient the common ancestor, the more losses are expected. The expected number of losses is proportional to the sum over all branch lengths in the species tree above the common ancestor in which losses may have occurred (in figure 4.1, all branches except those terminating in species C and D). The higher the number of losses per unit branch length, T , the high probability of one or more LGT events. For a protein family with L losses over a total branch length of B , the ratio T is simply:

$$T = L / B$$

Above some threshold minimum number of losses L_{\min} , the larger the value of T , the more likely there have been transfer events in the family. As discussed later, values of L_{\min} and T_{\min} were obtained by benchmarking against the most reliable predictions from the ERA method. This method is most sensitive for small families, with short total branch lengths.

High values of T and L identify families where transfer has taken place. To find the specific genes involved, we assume that the genes which require the most losses to accommodate by classical descent are the ones most likely to be the result of transfer. To identify these, each gene 'i' is removed from the family in turn, and the new minimum number of losses, L_i , calculated. The larger the difference ΔL_i between the numbers of losses required for the complete family, L_c , and L_i :

$$\Delta L_i = L_c - L_i$$

the more likely that gene 'i' is present as the result of an LGT event.

This approach can be extended to internal nodes of the tree, allowing the identification of some more ancient LGT events, occurring before species divergence. Each internal node with a value of '1', representing presence of an ancestral protein, is set to '0', and all its children nodes are also set to '0'. The reduction in the number losses in the family, ΔL , is then calculated as before. In some instances, the node resetting process may remove independent losses further towards the leaf nodes. To compensate for this effect, a modified reduction of losses, $\Delta L'$, is used for internal nodes:

$$\Delta L' = \Delta L - D$$

where the depth D is the number of steps needed to get to this node from a leaf node. D is the maximum number of losses that may be affected. As a consequence, $\Delta L'$ may be an underestimate of the loss reduction, and can lead to missing some more ancient events. A more sophisticated tree analysis might reduce this effect.

For each gene in a family, the maximum value of ΔL obtained by removing the gene or any of its ancestors is used. A threshold value ΔL_{\min} was obtained by benchmarking against the most reliable predictions from the ERA method.

The Evolutionary Rate Anomaly method (ERA)

Relative evolutionary rates of protein families.

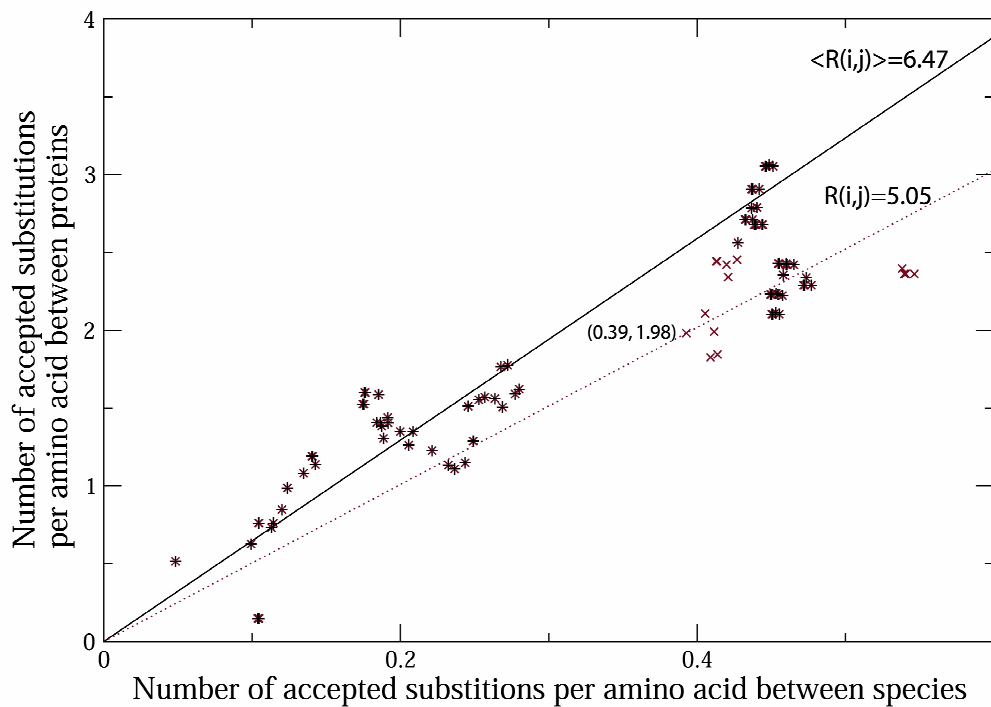


Figure 4.2. Example of Evolutionary Rate analysis of an Orthologous Protein Family (annotated as probable alkaline shock protein). The y value of each point is $S_f(i,j)$, the number of accepted substitutions per amino acid between proteins i and j in the orthologous family, and the x value is $S_{ref}(i,j)$, the average number of accepted substitutions in the 14 reference protein families between the species of proteins i and j . The relative evolutionary rate $R(i,j)$ is given by $S_f(i,j)/S_{ref}(i,j)$. The S values for one pair (GI:6459862 from *Deinococcus radiodurans* and GI:10175407 from *Bacillus halodurans*) are shown, corresponding to a relative rate of 5.05, the slope of the red dotted line. The relative evolutionary rate for the family, $\langle R(i,j) \rangle$ is derived from a robust linear regression technique – LMS (least median of squares), in this case 6.47, the slope of the black LMS line. Red crosses are points involving protein GI:6459862. A Student's

t-test shows the rates associated with this gene to be significantly lower than all the other rates, with 99.95% confidence, identifying a likely LGT event.

Rates of Amino Acid Substitution with Protein Families

Anomalous rates of change of amino acid sequences within an orthologous family are detected by comparing the number of accepted substitutions per amino acid between each pair of proteins i and j , $S_f(i,j)$, with the corresponding average number of accepted substitutions in the 14 reference protein families, $S_{ref}(i,j)$. The ratio of these quantities,

$$R(i,j) = S_f(i,j) / S_{ref}(i,j)$$

gives the relative rate of sequence change for the pair of proteins. For a protein family with an approximately constant rate of change throughout evolution, these ratios will be approximately the same for all protein pairs. A protein 'k' that has undergone a relatively recent lateral transfer will have anomalously low values of $R(k,j)$, with respect to all other proteins 'j'. Differences in these R values compared to all others can therefore be used to detect transfer events.

Figure 4.2 shows an example. A line from the origin through each point has a slope corresponding to the $R(i,j)$ value. The S values for one pair (GI:6459862 from *Deinococcus radiodurans* and GI:10175407 from *Bacillus halodurans*) are shown,

corresponding to a relative rate of 5.05, the slope of the red dotted line. The black line shows an estimate of the relative evolutionary rate for the whole family, $\langle R(i,j) \rangle$, derived from a robust linear regression technique – LMS (least median of squares), in this case 6.47. The red crosses show all the points involving *Deinococcus radiodurans*. All fall below the family rate line. A Student's t-test shows the rates for this protein to be significantly lower than all others, with high confidence (99.95%), strongly suggesting an LGT event.

Multiple transfer events will lead to more complex patterns of substitution relationships. In well behaved families, these can sometimes be resolved. For example, in figure 4.2, there is another gene (GI:15025074 from *Clostridium acetobutylicum*) whose rates are consistently lower than the family average. The Student's t-test indicates this gene is also transferred, with a confidence of 99.95%.

More ancient transfers, taking place before some speciation events, may result in a set of proteins with anomalous rates. This is most likely in regions where we have included a subset of genomes that have diverged relatively recently, for example, around *E.coli*. Particularly in smaller families, the present Student's t-test may not be able to resolve these. A more sophisticated approach, taking into account the tree structure, might do better.

Rates of substitution within families vary for other reasons besides lateral gene transfer. Apparently orthologous families may in fact contain proteins that have evolved different molecular functions, requiring substantial changes in sequence. Changes in the life style of a species, for example changed salinity or pH, may require adaptation of a protein sequence; changes in the pathways within an organism may impose new requirements on molecular function, also resulting in sequence adaptation. To help deal with these noisy data, we make use of LMS (Least Median of Squares), a robust linear regression technique, to retrieve the intrinsic relative evolutionary rate of the family $\langle R(i,j) \rangle$. However, the Student's t-test may deliver a less clear result when the underlying family clock is irregular.

Student's t-test for Anomalous Substitution Rates

A Student's t-test is performed for each protein 'k' in the family, evaluating the probability that the difference between the average rate for that protein is significantly different from the average rate over all other proteins i', considering the variances of the $[R(i',j)]$ and $[R(k,j)]$ distributions.

Points for which $S_{\text{ref}}(i,j) < 0.03$ are omitted from the $[R(i',j)]$ set, since S values were observed to be more noisy for small evolutionary distances. Outliers in the $[R(i',j)]$ distribution were also omitted, based on the value of

$$D = |R(i', j) - \langle R(i, j) \rangle|$$

For instance, in figure 4.2, a point close to the origin has a significantly different rate from $\langle R(i,j) \rangle$, and so should be omitted by the outlier filter. Benchmarking against the most confident set of HAGL LGT events was used to establish the optimum fraction of data points to omit, and the confidence threshold for the Student's t-test.

The Calculation of the number of accepted substitution per amino acid between proteins

Protein sequences in an orthologous family are aligned using ClustalW¹⁵⁸. Alignments were trimmed to include only those positions where at least 50% of the proteins in the family were aligned.

The Protdist module in the PHYLIP package (Felsenstein 1989)³¹ version 3.6 was used to compute a pair-wise protein distance matrix based on the truncated multiple sequence alignment. The amino acid substitution distance $S_f(i,j)$ between any pair of sequence i and j was derived from sequence identities, using the Jones-Taylor-Thornton matrix amino acid substitution model¹⁶³. The sequence identities are calculated from the aligned regions of a pair of two sequences, and gaps in the alignment are not considered.

LMS (Least Median of Squares)

The most popular linear regression technique is the Least Squares (least sum of squares) method. Given the data points $(x_1, y_1) \dots (x_n, y_n)$, the values 'a' and 'b' in the linear model

$$y = ax + b$$

are those which give the minimum of $\sum r_i^2$, where r_i is the residual of the i th data point, the difference between y_i and its estimated value y_i' , $y_i' = ax_i + b$. Least Squares is a simple and powerful method, but is extremely sensitive to outliers. Even one outlier may change the linear model significantly. The breakdown point (the smallest fraction of contamination that can falsify the linear estimator, where "falsify" is defined as changing the regression line by 90 degrees) of the Least Squares method is $1/n$, where n is the number of data.

As discussed above, relative substitution rate plots for protein families may have outliers because of lateral gene transfer events and other causes. To obtain reliable values of the relative substitution rate, we require a robust linear regression method (one with a breakdown point larger than $1/n$). LMS (Least Median of Squares) (Rousseeuw, P.J. 1984, J. Am. Stat. Assoc., 79, 871-880.)¹⁵⁷ is used. LMS finds a linear relationship which fits the majority of the data by minimizing the median of squares of residuals. That is, by choosing the line with:

$$\min(\text{med}[r_i^2])$$

where 'med' represents the median. The breakdown point of the method is 50%.

The $\text{med}[r_i^2]$ value is obtained for each line passing through the origin and a single data point (i.e. as many lines as data points). The slope of the line which minimizes the median of the squares of the residuals between the calculated and observed values of $S_f(i,j)$ (the number of accepted substitutions per site between family members in genomes 'i' and 'j') provides the estimate of the corresponding family's relative evolutionary rate $\langle R(i,j) \rangle$.

Calibration and Evaluation of the Methods

Accuracy Measures

The accuracy of the two methods is expressed in terms of specificity (fraction of true negatives correctly identified in a test set) and sensitivity (fraction of true positives correctly identified in a test set). I.e.

$$\text{Specificity (Sp)} = \text{TN} / (\text{TN} + \text{FP})$$

Where TN is the number of true negatives detected, FP is the number of false positives, and (TN + FP) is the total number of points in the set.

$$\text{Sensitivity (Sn)} = \text{TP} / (\text{TP} + \text{FN})$$

Where TP is the number of true positives detected, FN is the number of false positives, and (TP + FN) is the total number of points in the set.

Choice of Test Sets

Some putative lateral gene transfer events, such as those reported by ^{133; 142; 164} appear fairly certain, but at present, there is no way of compiling a highly reliable set of examples suitable for evaluating computational methods. Since we have developed two methods, we can partially evaluate them in terms of the agreement or otherwise in predicted LGT events. We take a subset of most reliable LGT events predicted by one method, and use it to obtain sensitivity data for the other. Similarly, a most reliable set of non-LGT events from one method is used to determine the specificity of the other. As noted earlier, the HAGL method performs best for small families, and the ERA method best for large families. The limited overlap of the methods restricts the size and reliability of the test sets. Low test set reliability generated by one method has the effect of causing the other to appear less accurate than it may be.

Optimization of Parameters

There are three parameters in The HAGL method (minimum number of losses (L), minimum rate of loss (T), and minimum reduction of losses on removing a candidate gene (ΔL)), and two parameters in the ERA method (confidence for the Student's t-Test, and fraction of distribution outliers excluded). Initial range estimates were made by inspection of the data, and each method was evaluated over those ranges. The most reliable subsets were then selected to provide the test data for final parameter choice. Merging of parameter calibration and method evaluation is not ideal, but unavoidable because of the limited test data. Since there are few parameters, and the sensitivity and specificity dependence are clear, it is a reasonable procedure in this case. In principle, this form of evaluation could lead to a false optimum in the parameter surface. That seems unlikely with these data, where specificity and sensitivity response to parameter change are straightforward.

4.3 Results

4.3.1 Phylogenetic Tree

Figure 4.3A shows the neighbor joining tree for the 66 genomes, built with distances derived from the average sequence identities over the set of 14 conserved protein families. Figure 4.3B shows the corresponding tree built with distances derived from 16S

rRNA sequence identities. The topologies of the trees are similar. The major kingdoms are well separated: Bacteria, Archaea and Eukaryotes, and each kingdom is further separated into small subgroups such as Proteobacteria, Actinobacteria and so forth. The tree for the 14 conserved protein families was used for the lateral gene transfer study. 16S ribosomal RNA, though well conserved across species, may be under different selection forces and so have different evolutionary rate properties from protein families.

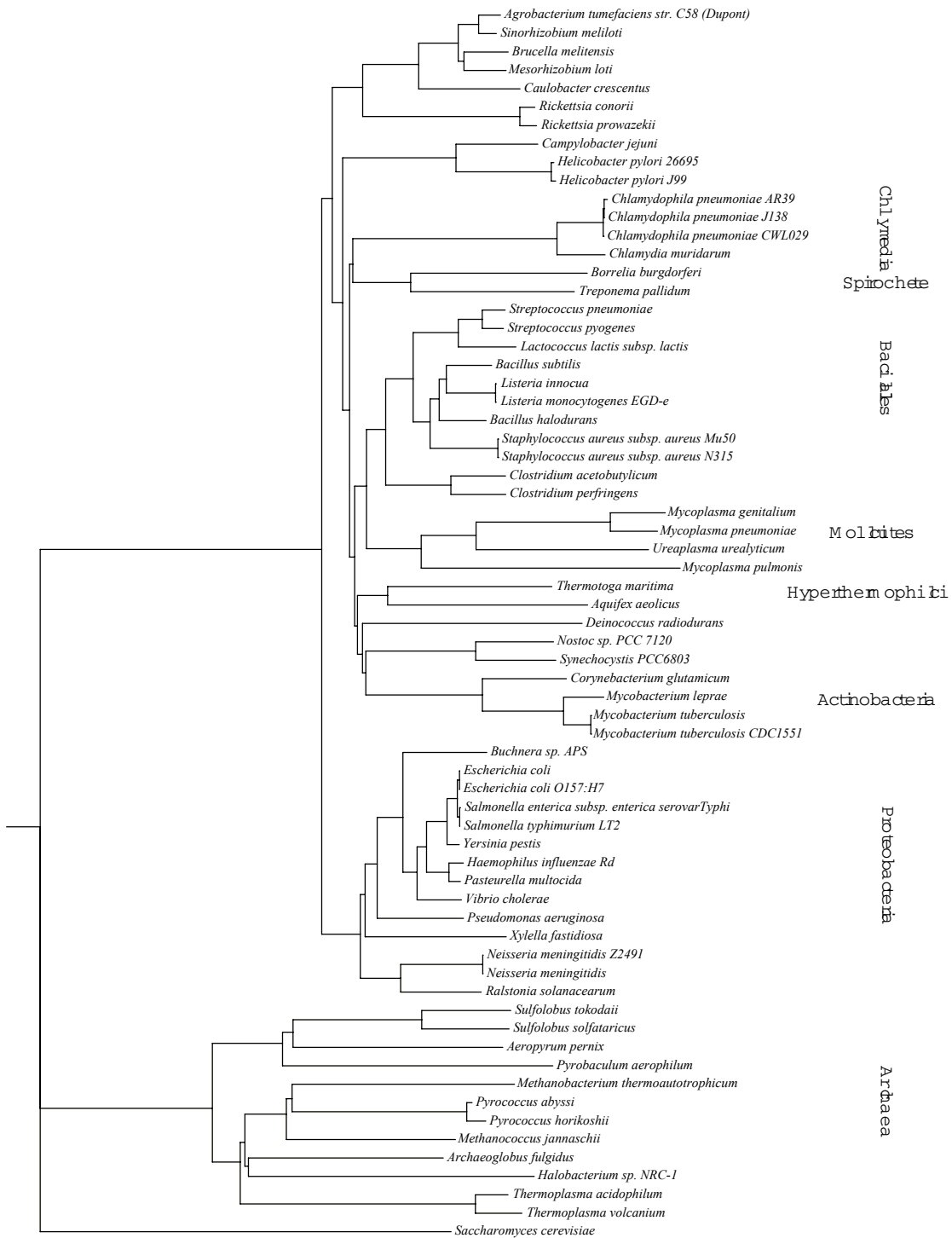


Figure 4.3A. The Neighbor Joining tree for 66 bacterial and Archaeal genomes, derived from 14 conserved protein families. *Saccharomyces cerevisiae* was used as an out-group.

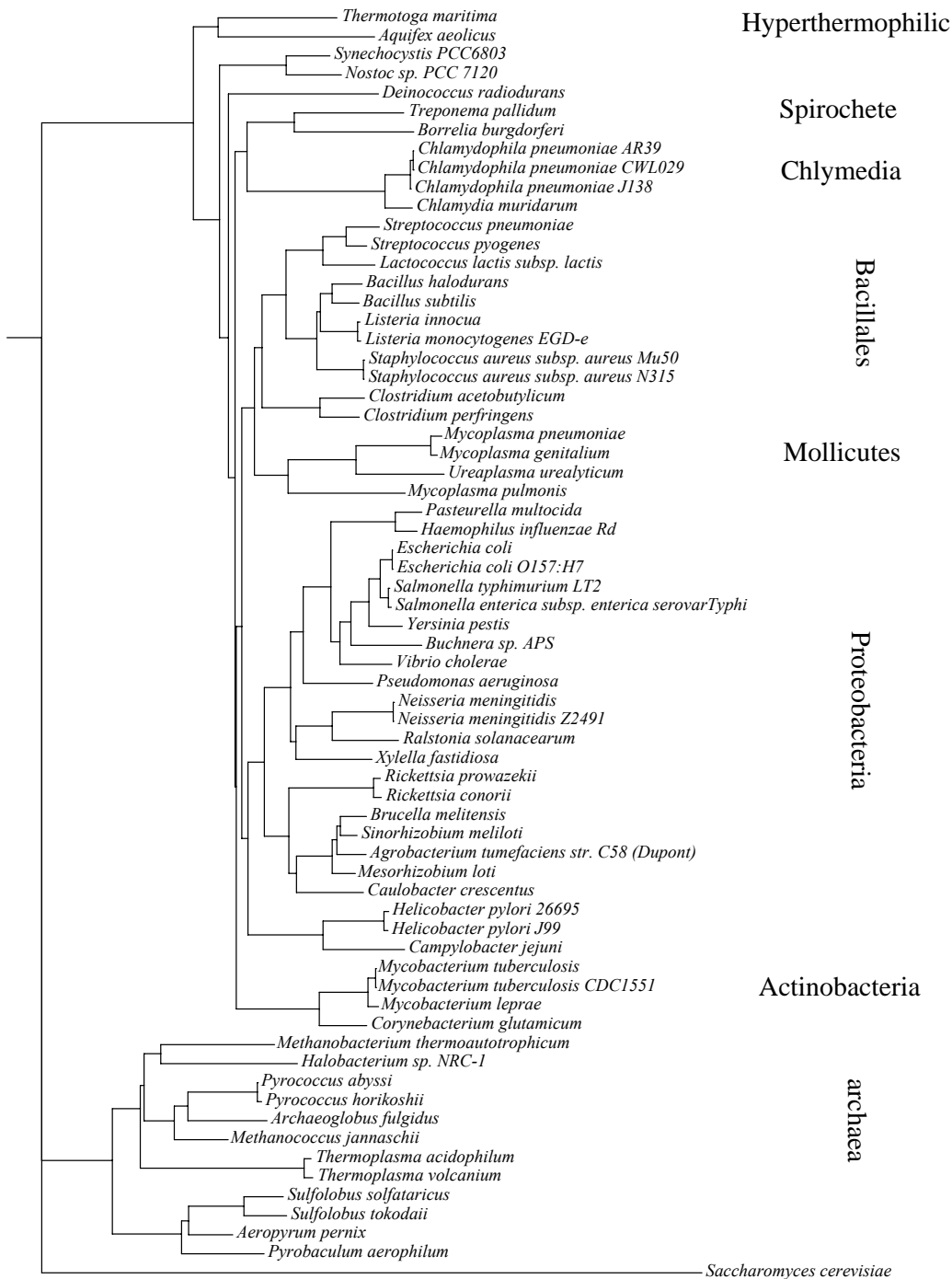


Figure 4.3B. 16S ribosomal RNA Neighbor Joining tree for the 66 genomes used in this study.

4.3.2 The High Apparent Gene Loss method (HAGL)

As discussed earlier, the HAGL method relates the likelihood of Lateral Gene Transfer in a protein family to the apparent number of gene loss events, as determined using Dollo Maximum Parsimony. The primary assumption is that there is a constant probability of a loss event per unit branch length in the phylogenetic tree of a family, so that families with a large ratio of losses to total branch length are the most likely to have experienced one or more LGT events.

Figure 4.4 shows the distribution of the number of orthologous families with two or more members, as a function of family size and number of apparent losses. There are a large number of small families (those with three to seven members) with many apparent losses (up to 15 losses is common). These are candidates for lateral gene transfer events, and are analyzed further. There is also a weaker concentration of families running along a diagonal line from top left to bottom right. This region represents ancient families with a relatively small number of losses. (Zero loss families present in most genomes have points high on the y axis. The more losses, the further down the diagonal). The HAGL method is not suitable for identifying LGT events in these families.

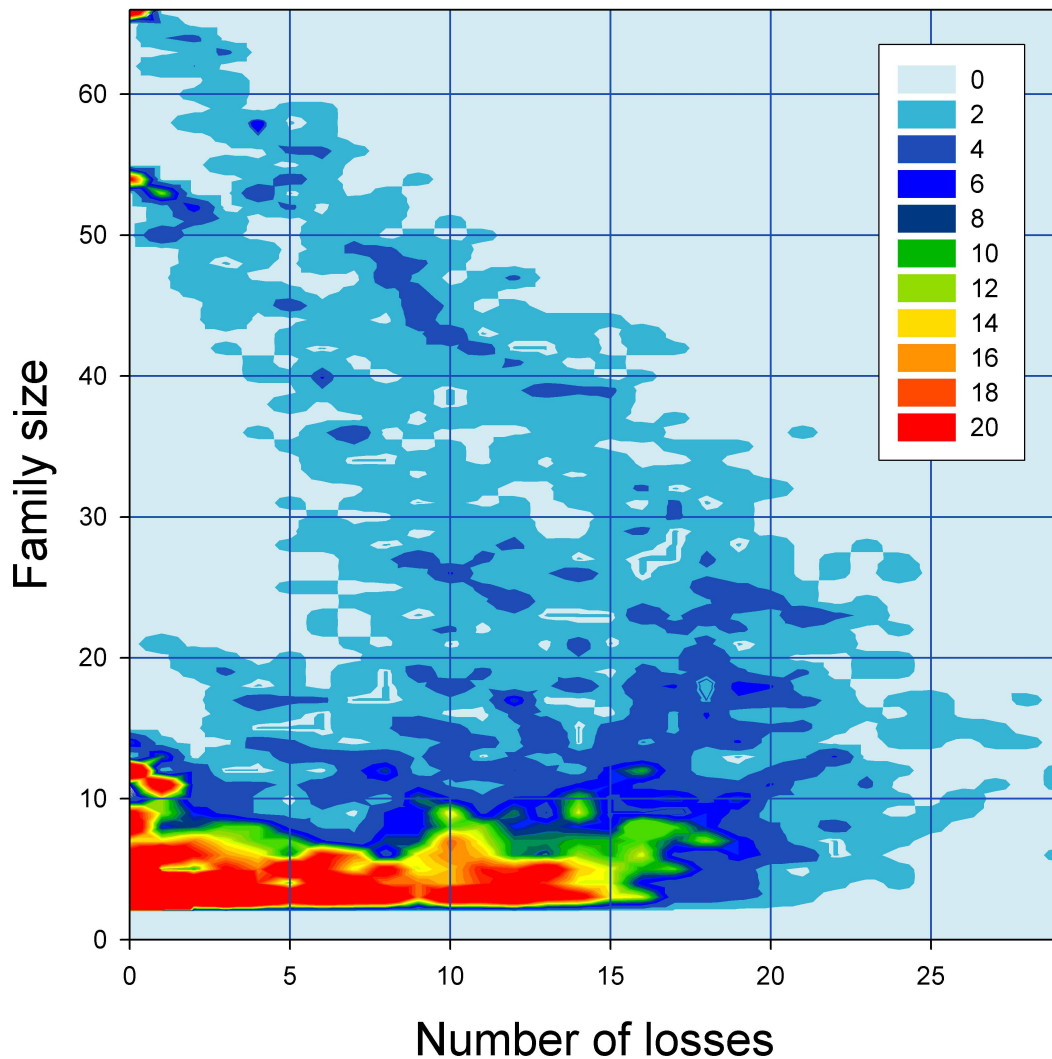


Figure 4.4: Distribution of protein families as a function of the number of apparent gene loss events in each family and family size. There are many small families with a large number of losses (represented by part of the red region running along the bottom right of the plot). Further analysis suggests many of these are not classical evolutionary descent families, but have undergone one or more lateral gene transfer events. The plot also shows a population of families along a diagonal line from top left towards the bottom

right. Features in this region represent classically descended families with a relatively small fraction of losses.

Figure 4.5 shows the distribution of the ratio of losses to total branch length, T , in these families. The large bar at the lowest T value represents the 1323 protein families which have T values less than 0.5. There is a tail of high T values. Benchmarking (described later) shows that value of T greater 5 is a reliable indicator of LGT, together with appropriate values of the other two parameters. Large T values may result from relatively few losses and very small branch lengths, for example in the *E.coli-Salmoneli* branch of the tree. We eliminate these from consideration by also requiring a minimum number of losses, established by benchmarking. For large families, there must be relatively few losses, and so T values tend to be small. As a result, the HAGL method is not suitable for detecting transfer events in these.

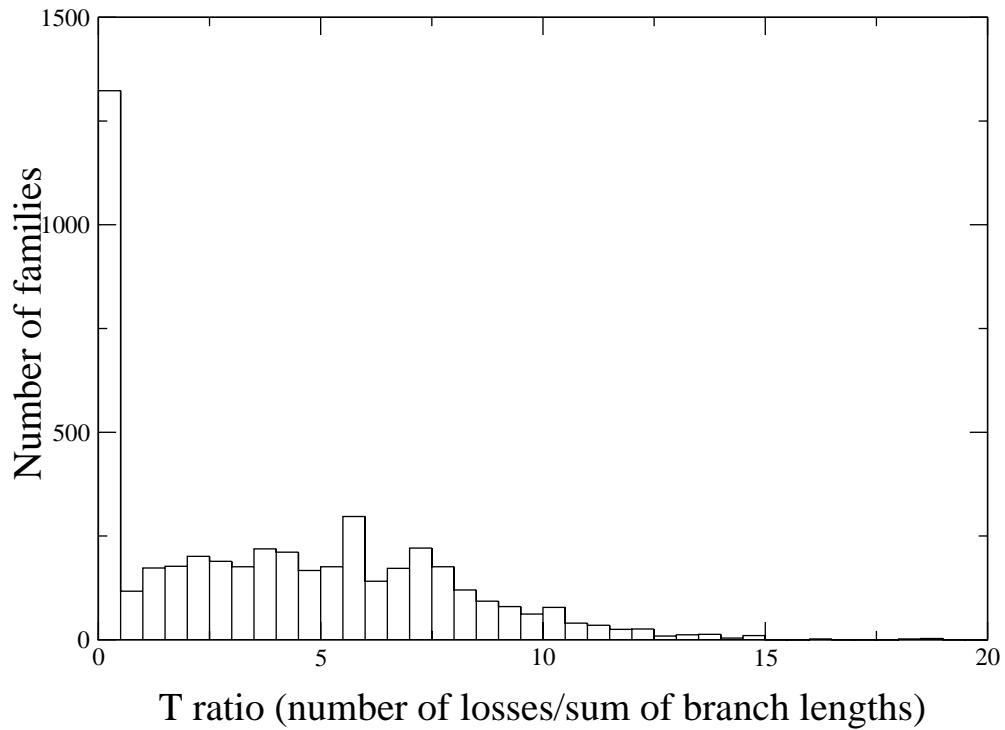


Figure 4.5. Distribution of T (ratio of losses to total branch length) for the 4856 orthologous protein families with three or more members. There are 1323 families with small T values (less than 0.5). The higher the value of T, the more likely are lateral gene transfer events.

4.3.3 The Evolutionary Rate Anomaly method (ERA)

As discussed earlier, the Evolutionary Rate Anomaly method (ERA) identifies LGT events by detecting genes that have a significantly different evolutionary rate from the rest of that family.

Figure 4.6A shows the distribution of average evolutionary rates for 4116 orthologous protein families, obtained using the LMS (Least Median of Squares) method. These rates are relative to that of the set of 14 highly conserved families. Most rates (98%) are between 1 and 10 times that of the reference families. 37 families with rates greater than 20 are not shown in the plot.

The distribution of standard deviations of evolutionary rates in families, $\sigma[(R(i,j) - \langle R(i,j) \rangle) / \langle R(i,j) \rangle]$, is shown in Figure 4.6B. The standard deviation is a measure of the irregularity of evolutionary rates within a family. A small standard deviation suggests the sequence changes in the corresponding family have occurred at a relatively constant rate throughout the history of the family. The larger the value, the more irregular the rate, and the less likely that the Student's t-test will be able to detect LGT events.

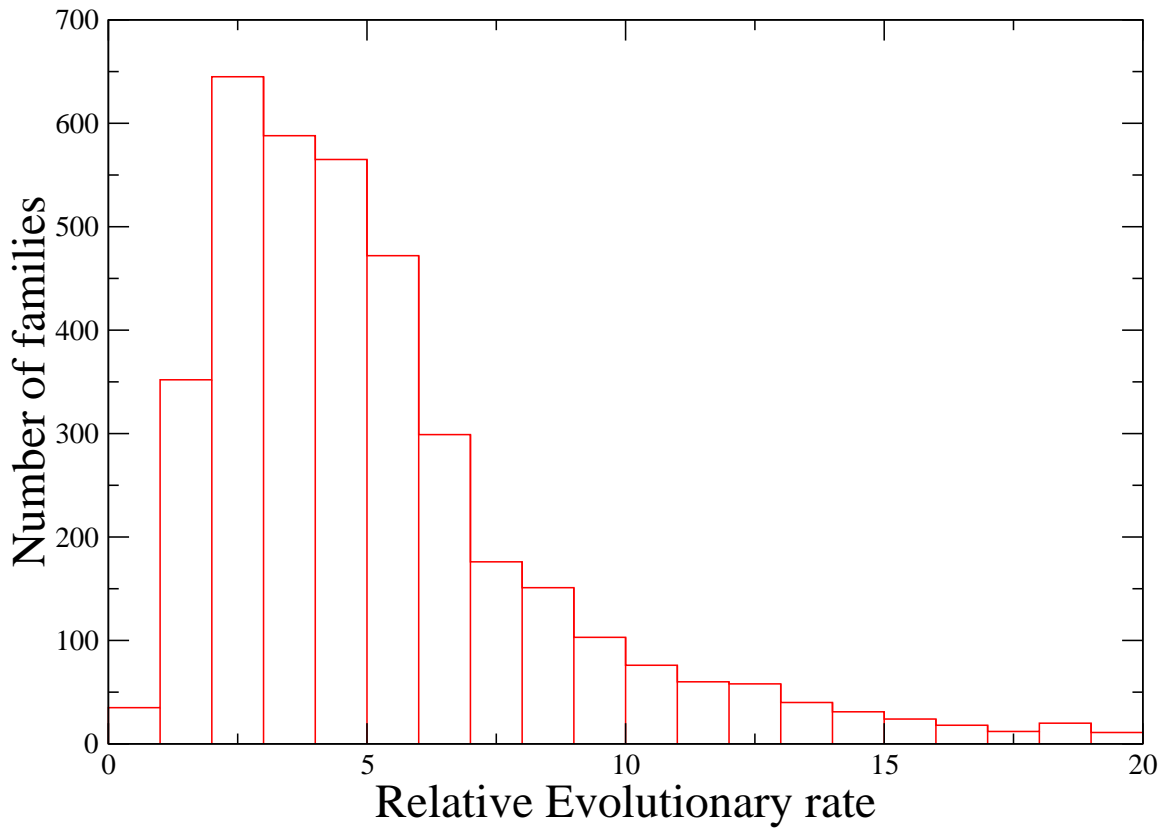


Figure 4.6A. Distribution of evolutionary rates for 4116 orthologous protein families. 98% of families have rates between 1 and 10 times that of the reference highly conserved families. There are 35 protein families with the rates less than 1, and 37 protein families with the rates larger than 20.

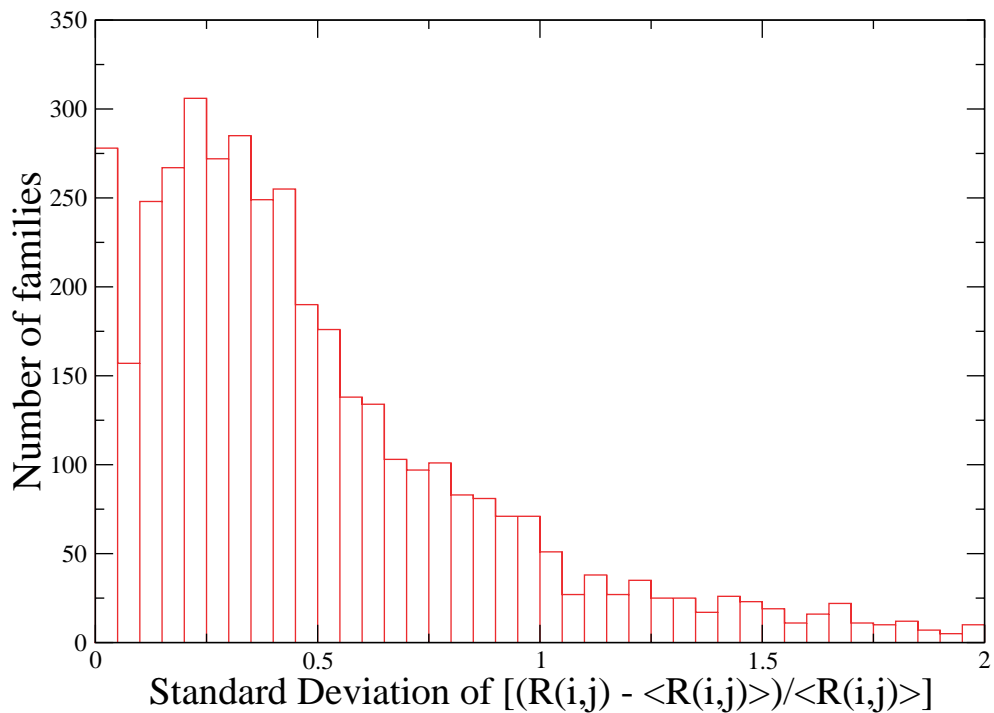


Figure 4.6B. Distribution of $\sigma[(R(i,j) - \langle R(i,j) \rangle) / \langle R(i,j) \rangle]$, the relative standard deviation of evolutionary rates within protein families. The smaller the value, the more constant the apparent rate of sequence change. A standard deviation of up to half the relative rate is common, and there is a tail of highly variable families.

Probable lateral gene transfer events were identified in the 2964 families with more than five members, using the Student's t-test to identify those proteins with anomalously low apparent evolutionary rates, as described in the methods section. The threshold for the t-test and the fraction of data included were determined by benchmarking against the most confident HAGL predictions, as described later.

An example of an identified LGT event is shown earlier in figure 4.2. A second example is shown in Figure 4.7A, for a family with 40 members, and an evolutionary rate of 5.3. Annotations suggest this family is a probable mercuric resistance operon repressor protein (*merR*). All the rates for Genbank ID 2649944 from *Archaeoglobus fulgidus* (shown as crosses) are anomalously low, and the Student's t-test gives a 99.95% probability that the rate for this protein is different from that of the rest of the family.

The HAGL method identifies a minimum of 10 losses in this family, over a total branch length of 5.39, giving a T value (ratio of the losses to the sum of branch lengths) of 1.86. With a T threshold of 5, this family is far from classification as involved in an LGT event, reflecting the insensitivity of the HAGL method with larger families.

Comparison of the topology of the phylogenetic tree of this family with the species tree (Figure 4.7B) shows the anomalous properties of this gene clearly. In the species tree, *Archaeoglobus fulgidus*, an archaean, is far from the bacteria. In the family tree, the gene from *Archaeoglobus fulgidus* is closer to the gene from *Bacillus subtilis* than the genes from other bacteria. This inconsistency suggests that gene 2649944 from *Archaeoglobus fulgidus* is very likely to have been lateral transferred relatively recently, after the divergence of *Bacillus subtilis* and *Streptococcus pyogenes*.

Function annotation for this family suggests it plays a regulatory role in mercury metabolism, and is one of a number of proteins needed for that function. In *Archaeoglobus fulgidus*, unlike most bacteria, only one other member of the pathway, the

periplasmic merP, is identifiable. The other components, such as the structural proteins merA and merB and the reductase merC, have not been found. So the functional role the transferred merR plays in *Archaeoglobus fulgidus* is an open question.

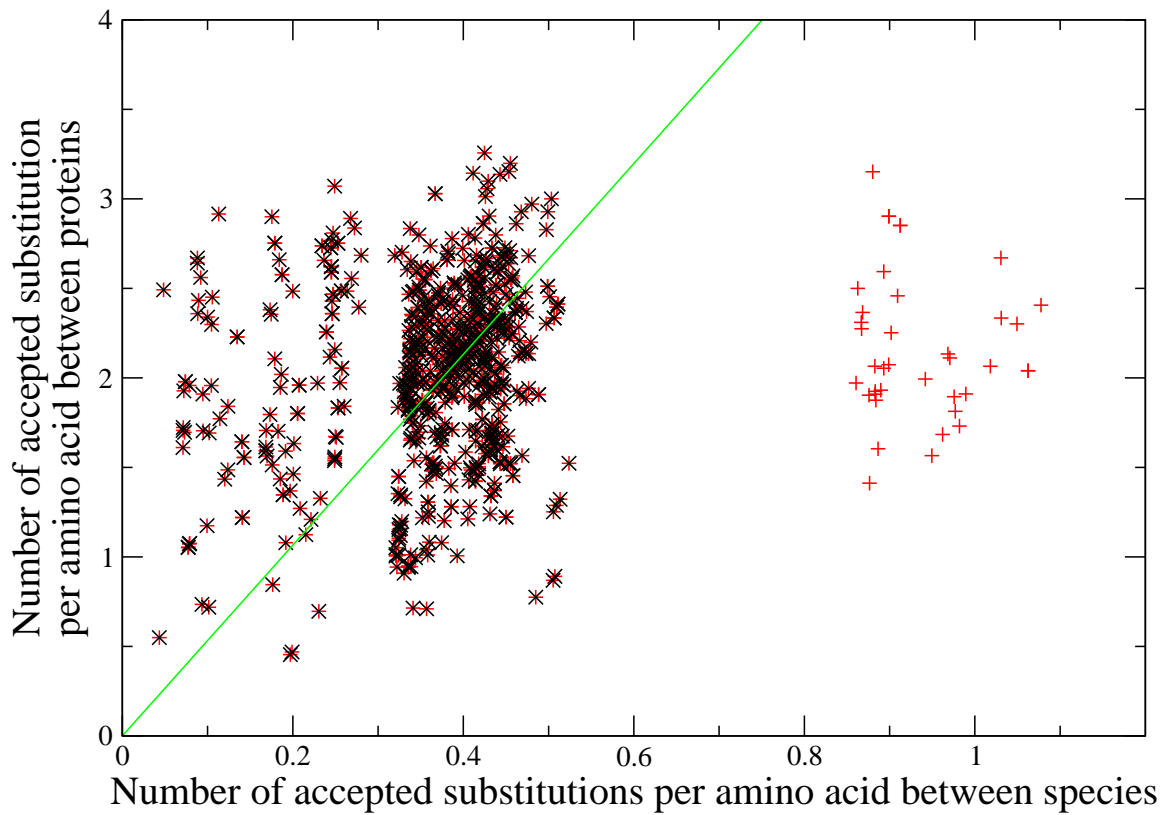


Figure 4.7A. Relative rates of amino acid substitution between pairs of proteins in the mercuric resistance operon regulatory protein family (merR). The vertical co-ordinate of each point is the number of accepted substitutions per residue between a pair of proteins

in the family, and the horizontal axis is the corresponding number of substitutions in a reference set of conserved families. The family evolutionary rate is estimated to be 5.3, the slope of the green LMS regression line. The set of rates ('+' points) for the protein from *Archaeoglobus fulgidus* are significantly different from the rest of family (99.95% confidence on a Student's t-test), indicating a lateral transfer event for this gene.

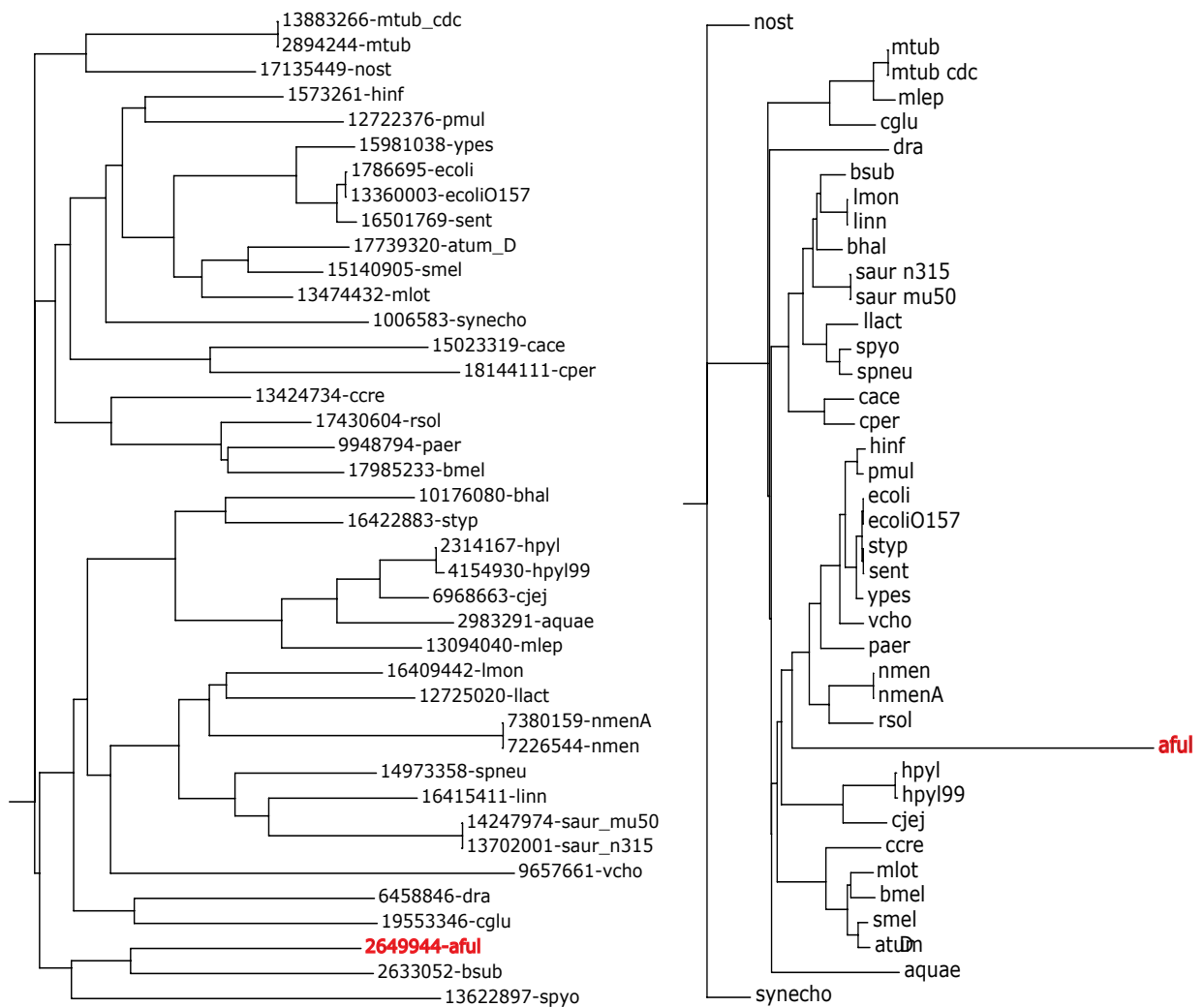


Figure 4.7B. The phylogenetic trees of the mercuric resistance operon repressor protein (merR) family (left) and the corresponding region of the species tree (right). In the species tree, *Archaeoglobus fulgidus* (aful), the only archaeal member of the family (shown in red color), is far from bacteria. In the family tree, the gene from this organism ('2649944-aful', shown in red color) is close to *Bacillus subtilis*. This topology difference suggests the gene 2649944 in *Archaeoglobus fulgidus* is very likely to have undergone lateral transfer.

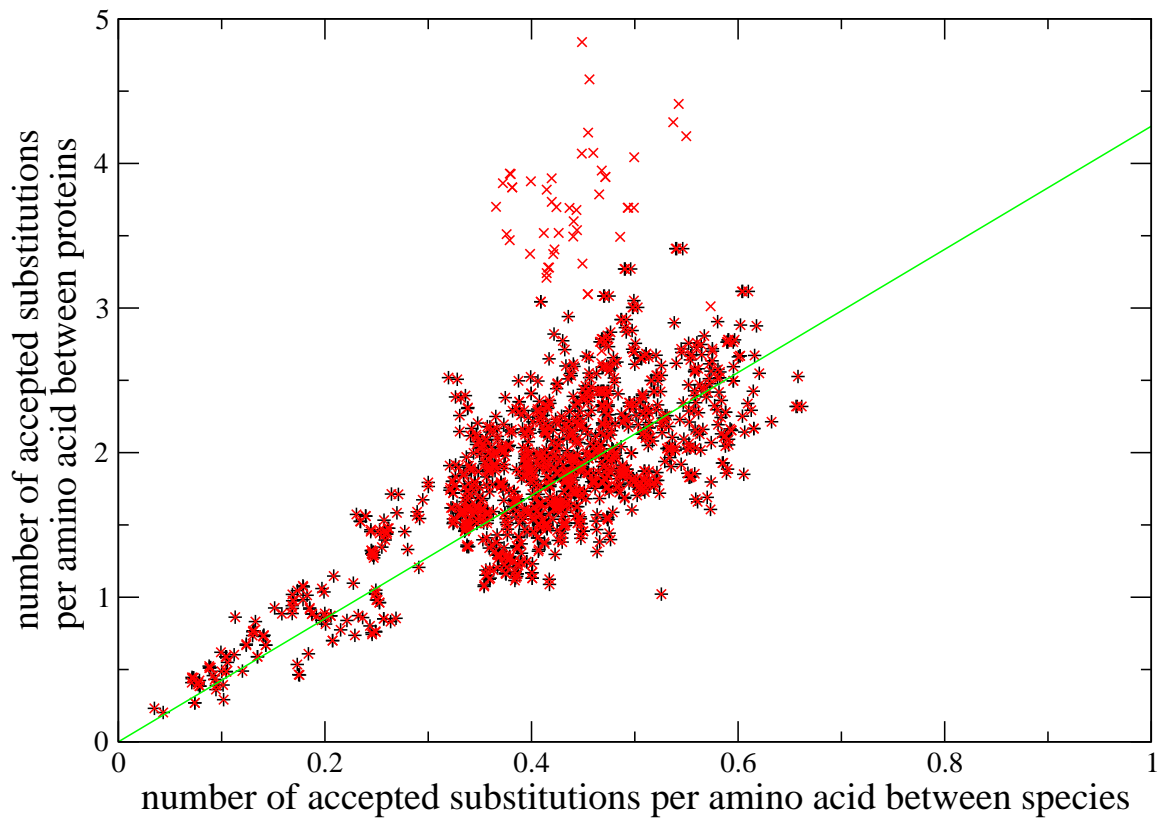


Figure 4.8A. Example of a protein family with anomalously high rates of amino acid change for one of its members. This protein family has 51 members, all in bacteria. These proteins may act as endonucleases in recombination. The vertical co-ordinate of each point is the number of accepted substitutions per residue between a pair of proteins in the family, and the horizontal axis is the corresponding number of substitutions in a reference set of conserved families. The family evolutionary rate is estimated to be 4.3, the slope of the green LMS regression line. The set of rates ('x' points) for the protein from gene 4982112 from *Thermotoga maritima* are significantly different from the rest (99.95% confidence on a Student's t-test), showing this gene has a faster evolutionary rate than the rest of the family.

ERA analysis may also identify proteins with anomalously rapid rates of sequence change. Figure 4.8A shows an example, where all the rates for the protein from *Thermotoga maritima* in a putative endonuclease protein family have high relative rates. This orthologous family has 51 members, all in bacterial species, with only three apparent losses. The Student's t-test returns a 99.95% confidence that the rates for this protein are significantly different from the others. Figure 4.8B shows that branch length for this protein in the family tree is anomalously long compared to the species tree.

The function of this protein remains unclear. A few members are annotated as "putative endonuclease involved in recombination and possible Holliday junction resolvase". The likely explanation for the anomalously high rate of sequence change in the *Thermotoga maritima* protein is that it has evolved to perform a different function. Examination of the

sequence alignment supports this suggestion. It has 218 amino acids whereas all other members of the family have lengths between 119 and 184, and there are two insertions in the middle of the sequence, as well as a locally weak alignment, all consistent with adaptation to a different function.

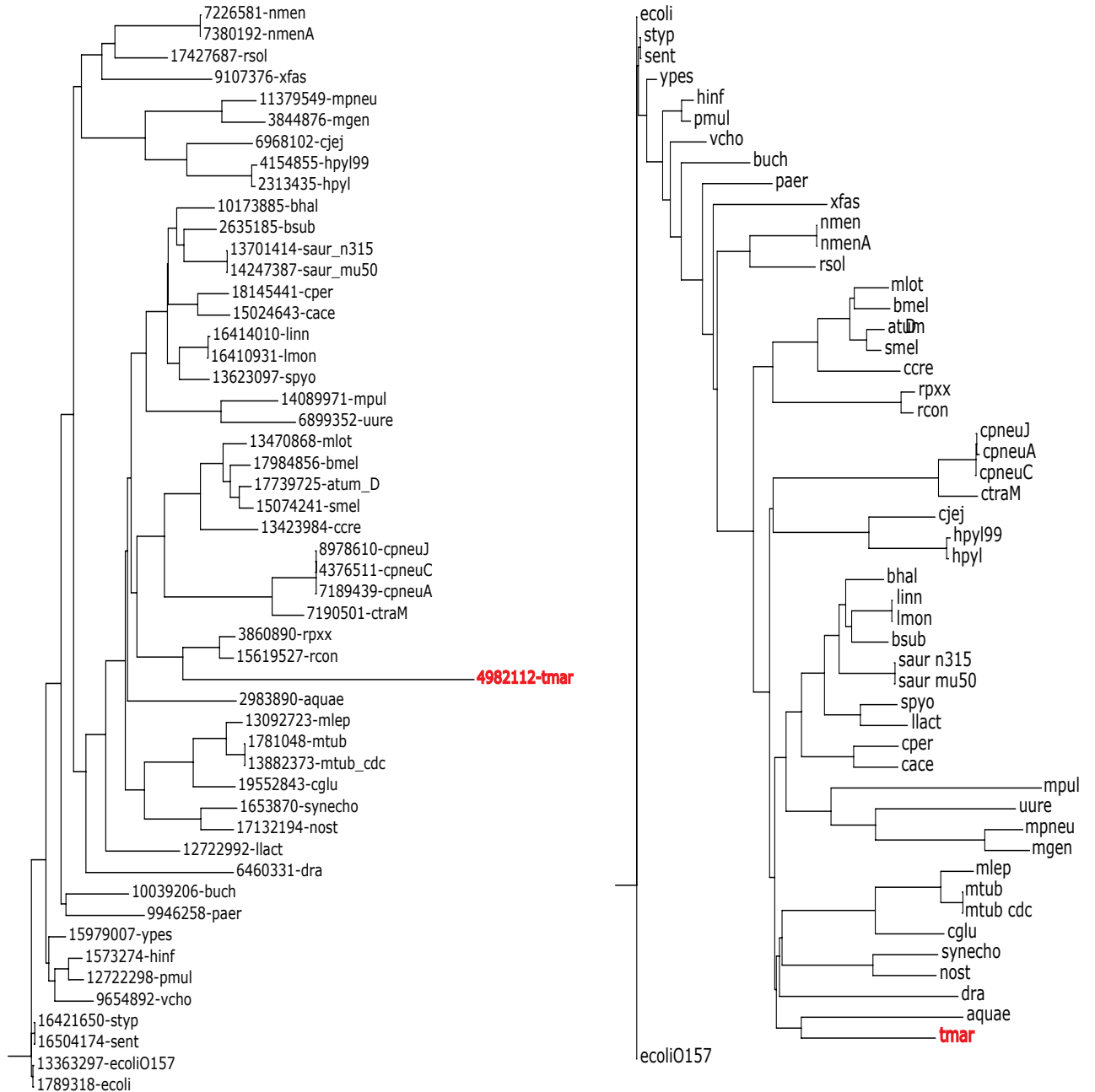


Figure 4.8B. The phylogenetic tree of a putative endonuclease protein family (left) and the corresponding portion of the species tree (right). This family has 51 members, all in bacteria. The protein 4982112 (shown in red) from *Thermotoga maritima* has an anomalously long branch length in the family tree, indicating rapid rate of sequence change, consistent with the ERA analysis.

4.3.4 Calibration and Evaluation of the Methods

As described earlier, each method was calibrated and evaluated against test sets of the most reliable results from the other. For the ERA method, the false positive rate is estimated using a set of 691 genes, all from families which have no apparent losses according to maximum parsimony, and with family sizes between 6 and 10. Any ERA assigned LGT events in this set are assumed to be false positives. The test set for the ERA false negative rate is 303 genes assigned as transferred by HAGL (rate of loss (T) larger than 5, at least 6 losses, a minimum reduction of losses of 4, family size between 6 and 10). Any of these genes not classified as transferred by ERA is counted as a false negative. Two parameters in the ERA method, the percentage of outlier points omitted and the P value for the Student's t-test, are optimized using this benchmarking scheme. The result is shown in table 4.3A: when 90% of data points used and the threshold of P value is 99.9%, 53% of LGT events are identified by ERA with a specificity of 95%. These parameter values were used for the genome wide analysis.

For the HAGL method, 2304 ERA non-LGT genes (Student's t-tests show these genes to have rates higher than the family average, family size between 10 and 15) were used to estimate the false positive rate. Any of these for which HAGL identifies LGT is counted as a false positive. A set of 529 genes classified as laterally transferred by the ERA method with high confidence ($> 99.95\%$), and omitting 10% of outliers, forms the false negative testing set. Any of these genes not assigned as laterally transferred by HAGL are counted as a false negative. The three HAGL parameters: the minimum number of losses per unit branch length, T_{\min} ; the minimum number of losses, L_{\min} ; and the reduction of losses ΔL on gene omission, were optimized with this benchmark. The results are shown in table 4.3B. With at least 6 losses, a T value larger than 5, and the minimum reduction of losses on gene removal at least 4, HAGL identifies 44% of LGT events with specificity of 96%. These parameter values were used for the genome wide analysis.

Percentage of data points for Student's t-test	Threshold of Student's t-test score	Number of genes correctly classified as LGT (out of 303)	Sensitivity	Number of genes correctly classified as Non-LGT (out of 691)	Specificity
100%	90%	172	56.8%	356	51.5%
	95%	167	55.1%	429	62.1%
	99%	162	53.4%	533	77.1%
	99.5%	149	49.2%	614	88.9%
	99.9%	134	44.2%	662	95.8%
	99.95%	125	41.3%	675	97.7%
90%	90%	207	68.3%	324	46.9%
	95%	205	67.7%	408	59.0%
	99%	191	63.0%	521	75.4%
	99.5%	176	58.1%	609	88.1%
	99.9%	160	52.8%	658	95.2%
	99.95%	141	46.5%	663	95.9%
75%	90%	209	69.0%	282	40.8%
	95%	207	68.3%	364	52.7%
	99%	199	65.7%	458	66.3%
	99.5%	188	62.0%	531	76.8%
	99.9%	166	54.8%	589	85.2%
	99.95%	149	49.2%	629	91.0%

Table 4.3A. Calibration and Evaluation of the ERA method. Sensitivity is measured by the fraction of 303 LGT events that are identified. Specificity is measured by the fraction of 691 genes with no LGT that are so classified. Both test sets are high confidence HAGL method assignments. Results for a range of Student's t-test thresholds are shown, as well as three thresholds for inclusion of rate outliers. On the basis of these data, a Student's t-test threshold of 99.9% and inclusion of 90% of the data points were chosen, yielding a sensitivity of 53% and a specificity of 95%.

Cutoff Parameters	Threshold	Number of genes correctly classified as LGT (out of 529)	Sensitivity	Number of genes correctly classified as Non-LGT (out of 2304)	Specificity
Minimum losses L_{\min}	0	237	44.8%	2145	93.1%
	2	237	44.8%	2187	94.9%
	4	236	44.6%	2202	95.6%
	6	235	44.4%	2221	96.4%
	8	231	43.7%	2228	96.7%
	10	220	41.6	2233	96.9%
Minimum rate of loss	3	311	58.8%	1734	75.3%
	4	275	52.0%	2024	87.8%
	5	235	44.4%	2221	96.4%

T	6	163	30.8%	2238	97.1%
	7	94	17.8%	2256	97.9%
Minimum loss change ΔL	0	264	49.9%	2101	91.1%
	2	248	46.9%	2169	94.1%
	4	235	44.4%	2221	96.4%
	6	172	32.5%	2260	98.1%
	8	40	7.6%	2292	99.4%

Table 4.3B Calibration and Evaluation of the HAGL method. Sensitivity is measured as the fraction of 237 LGT events identified. Specificity is measured by the fraction of 2304 genes with no LGT so classified. Test sets are high confidence ERA assignments. Results for a range of rates of apparent gene loss (T), number of gene losses (L), and reduction in gene loss on eliminating the candidate gene or genes (ΔL) are shown. On the basis of these data, thresholds of at least 6 losses, a T value (rate of loss) of at least 5 and a reduction in losses of at least 4 were selected, yielding a sensitivity of 44% and a specificity of 96%. Data for each parameter were obtained using the final values of the other two.

4.3.5 Application to the Set of 66 Genomes

Both methods were applied to all applicable proteins in the set of orthologous proteins, using the parameters derived in the previous section, and requiring at least

three family members for the HAGL method and at least six for the ERA method. Table 4.4 shows the number of LGT events identified in each genome by each method, and the total percentage of genes involved in LGT. As noted earlier, these numbers do not reflect all the LGT events in these genomes, and there may also be some method dependent biases. Nevertheless, interesting variations can be seen. The number of genes involved varies over a large range, from 3% of analyzed proteins in *Mycoplasma genitalium* and *Buchnera sp. APS*, to 33% in *Nostoc sp. PCC7120*, a Cyanobacterium and *Halobacterium sp. NRC-1*, an archaon. Organisms with small genomes, such as *Mycoplasma genitalium*, tend to have acquired fewer genes by recent LGT. This observation may be related to the fact that organisms with small genomes are mostly symbionts and also reflect constraints imposed by genome size limits. Analysis in terms of phylogenetic divisions suggests that some subgroups are more likely than others to accumulate LGT genes. For example, archaeal organisms generally have a higher percentage of assigned LGT genes: 18-33%. It is possible that this is a consequence of most archaea living in extreme conditions, such as high temperature or high pressure environments, and adaptation to these conditions is aided by the acquisition of new genes. It may also be that for some reason these organisms are more receptive to foreign genetic material.

Genome	Genome size	# of genes analyzed	Number of genes assigned LGT by ERA	Number of genes assigned LGT by HAGL	Total number of genes assigned LGT	%
<i>Aeropyrum pernix</i>	2694	552	86	49	102	18
<i>Agrobacterium tumefaciens str. C58 (Dupont)</i>	5402	1383	163	200	285	21
<i>Aquifex aeolicus</i>	1553	612	145	64	164	27
<i>Archaeoglobus fulgidus</i>	2407	712	124	131	202	28
<i>Bacillus halodurans</i>	4066	1170	106	192	236	20
<i>Bacillus subtilis</i>	4100	1179	107	164	210	18
<i>Borrelia burgdorferi</i>	1637	318	69	8	70	22
<i>Brucella melitensis</i>	3198	1141	125	111	173	15
<i>Buchnera sp. APS</i>	574	346	10	5	10	3
<i>Campylobacter jejuni</i>	1629	661	56	20	65	10
<i>Caulobacter crescentus</i>	3737	1102	185	98	229	21
<i>Chlamydia muridarum</i>	916	489	37	30	50	10
<i>Chlamydophila pneumoniae</i> AR39	1110	612	41	31	54	9
<i>Chlamydophila pneumoniae</i> CWL029	1052	614	38	31	51	8
<i>Chlamydophila pneumoniae</i> J138	1069	614	38	31	51	8
<i>Clostridium acetobutylicum</i>	3672	915	87	121	159	17
<i>Clostridium perfringens</i>	2723	885	108	102	155	18
<i>Corynebacterium glutamicum</i>	3040	870	106	108	168	19

<i>Deinococcus radiodurans</i>	3102	828	187	104	244	29
<i>Escherichia coli</i>	4289	1659	115	238	288	17
<i>Escherichia coli O157:H7</i>	5361	1742	117	279	331	19
<i>Haemophilus influenzae Rd</i>	1709	823	51	85	103	13
<i>Halobacterium sp. NRC-1</i>	2605	616	150	128	204	33
<i>Helicobacter pylori 26695</i>	1566	571	44	54	70	12
<i>Helicobacter pylori J99</i>	1490	569	47	54	73	13
<i>Lactococcus lactis subsp. lactis</i>	2266	727	63	67	88	12
<i>Listeria innocua</i>	3043	1027	99	135	179	17
<i>Listeria monocytogenes EGD-e</i>	2846	1013	96	122	164	16
<i>Mesorhizobium loti</i>	7275	1538	196	304	399	26
<i>Methanobacterium thermoautotrophicum</i>	1869	603	118	87	149	25
<i>Methanococcus jannaschii</i>	1770	606	94	75	120	20
<i>Mycobacterium leprae</i>	1605	756	57	88	107	14
<i>Mycobacterium tuberculosis</i>	3869	1119	170	176	244	22
<i>Mycobacterium tuberculosis CDC1551</i>	4187	1112	170	175	243	22
<i>Mycoplasma genitalium</i>	480	219	6	3	6	3
<i>Mycoplasma pneumoniae</i>	688	231	11	9	14	6
<i>Mycoplasma pulmonis</i>	782	243	23	6	24	10
<i>Neisseria meningitidis</i>	2025	795	59	64	81	10
<i>Neisseria meningitidis Z2491</i>	2032	808	59	68	84	10
<i>Nostoc sp. PCC 7120</i>	6129	973	144	255	323	33
<i>Pasteurella multocida</i>	2014	881	73	86	115	13

<i>Pseudomonas aeruginosa</i>	5565	1398	164	182	248	18
<i>Pyrobaculum aerophilum</i>	2605	594	102	89	144	24
<i>Pyrococcus abyssi</i>	1765	667	121	108	173	26
<i>Pyrococcus horikoshii</i>	2064	642	112	112	157	24
<i>Ralstonia solanacearum</i>	5116	1300	177	185	259	20
<i>Rickettsia conorii</i>	1374	442	52	25	62	14
<i>Rickettsia prowazekii</i>	834	384	19	12	22	6
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovarTyphi	4749	1696	114	251	287	17
<i>Salmonella typhimurium</i> LT2	4553	1728	116	258	295	17
<i>Sinorhizobium meliloti</i>	6205	1496	173	260	321	21
<i>Staphylococcus aureus</i> subsp. <i>aureus</i> Mu50	2748	930	89	82	115	12
<i>Staphylococcus aureus</i> subsp. <i>aureus</i> N315	2624	914	89	72	107	12
<i>Streptococcus pneumoniae</i>	2094	690	63	56	77	11
<i>Streptococcus pyogenes</i>	1696	658	64	64	85	13
<i>Sulfolobus solfataricus</i>	2977	707	121	105	149	21
<i>Sulfolobus tokodaii</i>	2826	700	125	94	144	21
<i>Synechocystis</i> PCC6803	3169	785	122	134	190	24
<i>Thermoplasma acidophilum</i>	1478	521	87	69	110	21
<i>Thermoplasma volcanium</i>	1526	527	89	64	107	20
<i>Thermotoga maritima</i>	1846	688	126	113	183	27
<i>Treponema pallidum</i>	1031	327	38	24	45	14
<i>Ureaplasma urealyticum</i>	611	218	8	8	9	4
<i>Vibrio cholerae</i>	3828	1171	85	190	208	18
<i>Xylella fastidiosa</i>	2831	818	63	77	94	11

<i>Yersinia pestis</i>	4039	1392	100	164	197	14
------------------------	------	------	-----	-----	-----	----

Table 4.4. Number of analyzed genes in each of the 66 genomes that are assigned LGT events by each method, and the total percentage of these genes affected in each genome. These numbers underestimate the total amount of LGT that has taken place. The fraction of genes varies over a wide range, from 3% of analyzed proteins in *Mycoplasma genitalium* and *Buchnera sp. APS*, to 33% in *Nostoc sp. PCC7120*, a Cyanobacterium and *Halobacterium sp. NRC-1*, an archaon. The second column shows the number of genes in each organism, and the third, the number analyzed (those in orthologous families with three or more members).

4.4 Conclusion and Discussion

Two new approaches for studying lateral gene transfer have been developed: the High Apparent Gene Loss (HAGL) method and the Evolutionary Rate Anomaly (ERA) method. The HAGL method identifies LGT events by counting the minimum number of losses needed to explain the phyletic pattern of a protein family in terms of classical evolutionary descent. The higher the number of apparent losses and the smaller the evolutionary distance over which they occurred, the more likely that one or more lateral gene transfer events has taken place in the family. The specific genes involved are then identified by considering the reduction in the number losses in the family when each gene or sub-tree is removed. This method works best with small protein families with a large number of apparent losses. The other method, ERA,

detects LGT events by finding proteins with significantly slower apparent rates of sequence change than the rest of the family. Evolutionary rates within many families vary considerably for reasons other than lateral gene transfer. We reduce the impact of this noise by using a robust linear regression technique to find average rates for a family. A conventional Student's t-test is used to measure the probability that rates for one protein differ significantly from those of the rest of the family. Although uneven evolutionary rates limit application of the method, it is able to reliably detect a substantial number of LGT events. The method works best for larger families, where many genes determine the average evolutionary rate, and there are many rates involving each individual gene. The two methods are complementary, since HAGL works best with smaller families. Together, they detect a large number of LGT events, but by no means all. In addition to limitations imposed by noisy data, both methods require that LGT has taken place from an origin in near to a sequenced genome, and that the events be relatively recent. Nevertheless, sampling of LGT is sufficiently broad that many interesting cases are revealed, and an overall pattern in different genomes can be seen.

Evaluation of the methods is complicated by the absence of a known high reliability set of lateral gene transfer events. Some well known biologically reasonable cases, such as the transfer of the pathogenicity island into *E.coli O157*^{131; 133; 146; 164}, cannot be used because the origin of the transfer is so far unknown. We have benchmarked by using a more reliable subset of LGT assignments from each method as a test set for the other. Both methods yield high specificity (less than 10% false positives)

under conditions of moderate sensitivity (detecting about half of the LGT cases in the test set). The low sensitivity partly reflects the fact that the two methods work best under different circumstances – HAGL for small families and ERA for larger ones. For the ERA method, sensitivity is also limited by the effect of varying evolutionary rates within the classical evolutionary descent regions of families, some times making confident identification of abnormal rates difficult. As discussed below, there is a wide variation in the uniformity of evolutionary rates within families. For the HAGL method, a high apparent rate of gene loss is necessary for a confident prediction, reducing sensitivity.

The Least Median Squares analysis provides a set of relative evolutionary rates for protein families (Figure 4.6A). 98% of families have rates between 1 and 10 times that of the conserved reference set, with the most common value about 2.5 times reference. 35 families have rates slower than the reference value. Proteins involved in many interactions with other molecules, such as ribosome components, typically have small rates, while monomeric proteins with ‘simple’ functions in general have higher rates. So far, there has been no systematic study of these relationships, though. The new data provide a basis for such an analysis. The ERA method also detects anomalously fast rates of change of particular proteins. These may arise from imperfect extraction of orthologous subfamilies, but also from the result of gene fusion events (which tend to be associated with rapid sequence change), or changes in the network environment within a particular species. We have observed very high rates for some small families, such as the “Gifsy-1 prophage protein, a family with

four members (*E.coli K12*, *E.coli O157*, *Salmonella typhimurium* and *Salmonella enterica*). This phage family may be under very high selective pressure and also subject to much higher rates of accepted substitutions than those in prokaryotes. As figure 4.6B shows, families also differ widely in the consistency of the rates of divergence over the phylogenetic tree. Some of the irregularity comes from LGT events, and some from the accidental inclusion of proteins with different functions. A number of other causes are possible, such as the effects of gene fusion, and adaptation to changes in network environment. Again, there has so far been no systematic study of these factors, and the new data provide a basis for that.

A number of other LGT identification methods have been developed. Most of these are not easily scaled for the analysis of many genomes¹⁶⁵. One class of methods that can be applied on a genome scale are those that rely on identifying irregularities in gene composition. Garcia-Vallve et al suggested a method for detecting lateral gene transfer in terms of irregular GC content¹⁵¹. They consider genes as extraneous when their GC content deviates by $> 1.5\sigma$ from the genome mean value or when the GC content in the first and third codon positions deviate from the genome mean in the same direction and at least one of them is $> 1.5\sigma$. We implemented this method and found some unexpected results. For instance, the family of ribosomal protein S17 is an apparently strictly inherited protein family, with member in every species. 10 genes out of 66 members in this family are assigned as having undergone LGT, which seems very unlikely. Indications of over-prediction of LGT are in agreement with other studies^{153 141} which concluded that gene composition methods have low

reliability. These methods do have the advantage that, unlike HAGL and ERA, they can detect transfer events from outside currently unsampled genome space, provided the events are fairly recent.

There are a number of ways in which LGT detection methods may be improved in future. High confidence predictions from the HAGL and ERA methods may provide a test set for more reliable parameterization and testing of gene composition methods. It may be possible to assemble a large enough set of biologically reasonable cases to provide independent testing. Proper combining of available methods using a Bayesian approach or machine learning, such as a Support Vector Machine, would then make maximum use of the different signals. Increasing numbers of fully sequenced genomes will reduce the number of cases where the origin of a transfer cannot be identified, so increasing the applicability of the HAGL and ERA methods.

All family rates and variances, and LGT events identified by the HAGL and ERA methods are available at <http://moult.umbi.umd.edu/LGT/>.

Chapter 5 Conclusions

The overall conclusions are as follows:

- 1) I have developed two methods, the Gene Neighbor Method (GNM) and the Gene Gap Method (GGM), to predict operon structure in microbial genomes. The two methods were benchmarked with function pathway data and documented operon data. The primary use of the predictions is to infer the function of hypothetical proteins in genomes.
- 2) I have developed a protein family clustering procedure and successfully classified the proteins in a set of microbial genomes. This set of protein families is complete in terms of classifying all protein sequences. Benchmarking using SCOP data and PFAM data shows that this protein family set is more sensitive than sequence alignment methods, at a low false positive rate.
- 3) The protein family set was used to address several important questions in structural genomics: (1) What is the structure coverage for currently known families? (2) How will the number of known apparent families grow as more genomes are sequenced? (3) What is a practical strategy for maximizing structure coverage in future? Our study indicates that approximately 20% of known families with three or more members currently have a representative structure. The number of apparent protein families will be considerably larger than previously thought: We estimate that, by the criteria of this work, there

will be about 250,000 protein families when 1000 microbial genomes have been sequenced. However, the vast majority of these families will be small. It will be possible to obtain structural templates for 70 – 80% of protein domains with an achievable number of representative structures, by systematically sampling the larger families.

- 4) Two methods were developed to identify lateral gene transfer events in microbial genomes, the High Apparent Gene Loss method (HAGL) and the Evolutionary Rate Anomaly (ERA) method. Although the methods do not provide complete detection of all LGT events, together, they do give a useful sampling, and reveal considerable variance in the extent of LGT in different organisms.

Bibliography

1. Gabaldon, T. & Huynen, M. A. (2004). Prediction of protein function and pathways in the genome era. *Cell Mol Life Sci* 61, 930-44.
2. Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* 32 Database issue, D226-9.
3. Hubbard, T. J., Murzin, A. G., Brenner, S. E. & Chothia, C. (1997). SCOP: a structural classification of proteins database. *Nucleic Acids Res* 25, 236-9.
4. Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A. & Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* 26, 320-2.
5. Siew, N. & Fischer, D. (2003). Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins* 53, 241-51.
6. Siew, N. & Fischer, D. (2003). Twenty thousand ORFan microbial protein families for the biologist? *Structure (Camb)* 11, 7-9.
7. Vitkup, D., Melamud, E., Moul, J. & Sander, C. (2001). Completeness in structural genomics. *Nat Struct Biol* 8, 559-66.
8. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. *Structure* 5, 1093-108.
9. Davies, J. (1996). Origins and evolution of antibiotic resistance. *Microbiologia* 12, 9-16.
10. Huynen, M., Snel, B., Lathe, W., 3rd & Bork, P. (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res* 10, 1204-10.
11. Blumenthal, T., Evans, D., Link, C. D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W. L., Duke, K., Kiraly, M. & Kim, S. K. (2002). A global analysis of *Caenorhabditis elegans* operons. *Nature* 417, 851-4.
12. Spieth, J., Brooke, G., Kuersten, S., Lea, K. & Blumenthal, T. (1993). Operons in *C. elegans*: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. *Cell* 73, 521-32.
13. Yanofsky, C. & Lennox, E. S. (1959). Transduction and recombination study of linkage relationships among the genes controlling tryptophan synthesis in *Escherichia coli*. *Virology* 8, 425-47.
14. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* 96, 2896-901.
15. Lawrence, J. G. & Roth, J. R. (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* 143, 1843-60.

16. Yada, T., Nakao, M., Totoki, Y. & Nakai, K. (1999). Modeling and predicting transcriptional units of Escherichia coli genes using hidden Markov models. *Bioinformatics* 15, 987-93.
17. Thieffry, D., Salgado, H., Huerta, A. M. & Collado-Vides, J. (1998). Prediction of transcriptional regulatory sites in the complete genome sequence of Escherichia coli K-12. *Bioinformatics* 14, 391-400.
18. Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. & Shao, Y. (1997). The complete genome sequence of Escherichia coli K-12. *Science* 277, 1453-74.
19. Craven, M., Page, D., Shavlik, J., Bockhorst, J. & Glasner, J. (2000). A probabilistic learning approach to whole-genome operon prediction. *Proc Int Conf Intell Syst Mol Biol* 8, 116-27.
20. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. (1999). Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol* 1, 93-108.
21. Ermolaeva, M. D., White, O. & Salzberg, S. L. (2001). Prediction of operons in microbial genomes. *Nucleic Acids Res* 29, 1216-21.
22. Wolf, Y. I., Rogozin, I. B., Kondrashov, A. S. & Koonin, E. V. (2001). Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res* 11, 356-72.
23. Dandekar, T., Snel, B., Huynen, M. & Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 23, 324-8.
24. Tamames, J. (2001). Evolution of gene order conservation in prokaryotes. *Genome Biol* 2, 0020.
25. Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L. & Koonin, E. V. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol Biol* 1, 8.
26. Salgado, H., Moreno-Hagelsieb, G., Smith, T. F. & Collado-Vides, J. (2000). Operons in Escherichia coli: genomic analyses and predictions. *Proc Natl Acad Sci U S A* 97, 6652-7.
27. Kanehisa, M. (2002). The KEGG database. *Novartis Found Symp* 247, 91-101; discussion 101-3, 119-28, 244-52.
28. Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Millan-Zarate, D., Diaz-Peredo, E., Sanchez-Solano, F., Perez-Rueda, E., Bonavides-Martinez, C. & Collado-Vides, J. (2001). RegulonDB (version 3.2): transcriptional regulation and operon organization in Escherichia coli K-12. *Nucleic Acids Res* 29, 72-4.
29. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-402.
30. Altschul, S. F. & Koonin, E. V. (1998). Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. *Trends Biochem Sci* 23, 444-7.

31. Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4, 406-25.
32. Woese, C. R. (1987). Bacterial evolution. *Microbiol Rev* 51, 221-71.
33. Wixon, J. & Kell, D. (2000). The Kyoto encyclopedia of genes and genomes--KEGG. *Yeast* 17, 48-55.
34. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. & Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 27, 29-34.
35. Kanehisa, M. & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28, 27-30.
36. Akiyama, M., Crooke, E. & Kornberg, A. (1993). An exopolyphosphatase of *Escherichia coli*. The enzyme and its ppx gene in a polyphosphate operon. *J Biol Chem* 268, 633- 9.
37. Bae, W., Xia, B., Inouye, M. & Severinov, K. (2000). *Escherichia coli* CspA-family RNA chaperones are transcription antiterminators. *Proc Natl Acad Sci U S A* 97, 7784-9.
38. Charpentier, B., Bardey, V., Robas, N. & Branlant, C. (1998). The EIIGlc protein is involved in glucose-mediated activation of *Escherichia coli* gapA and gapB-pgk transcription. *J Bacteriol* 180, 6476-83.
39. Clark, M. A., Baumann, L. & Baumann, P. (1992). Sequence analysis of an aphid endosymbiont DNA fragment containing rpoB (beta-subunit of RNA polymerase) and portions of rplL and rpoC. *Curr Microbiol* 25, 283-90.
40. Coleman, J. & Raetz, C. R. (1988). First committed step of lipid A biosynthesis in *Escherichia coli*: sequence of the lpxA gene. *J Bacteriol* 170, 1268-74.
41. Egan, S. E., Fliege, R., Tong, S., Shibata, A., Wolf, R. E., Jr. & Conway, T. (1992). Molecular characterization of the Entner-Doudoroff pathway in *Escherichia coli*: sequence analysis and localization of promoters for the edd-eda operon. *J Bacteriol* 174, 4638-46.
42. Hesslinger, C., Fairhurst, S. A. & Sawers, G. (1998). Novel keto acid formate-lyase and propionate kinase enzymes are components of an anaerobic pathway in *Escherichia coli* that degrades L-threonine to propionate. *Mol Microbiol* 27, 477-92.
43. Jerlstrom, P. G., Bezjak, D. A., Jennings, M. P. & Beacham, I. R. (1989). Structure and expression in *Escherichia coli* K-12 of the L-asparaginase I-encoding ansA gene and its flanking regions. *Gene* 78, 37-46.
44. Jordan, A., Pontis, E., Aslund, F., Hellman, U., Gibert, I. & Reichard, P. (1996). The ribonucleotide reductase system of *Lactococcus lactis*. Characterization of an NrdEF enzyme and a new electron transport protein. *J Biol Chem* 271, 8779 -85.
45. Kamio, Y., Lin, C. K., Regue, M. & Wu, H. C. (1985). Characterization of the ileS-lsp operon in *Escherichia coli*. Identification of an open reading frame upstream of the ileS gene and potential promoter(s) for the ileS-lsp operon. *J Biol Chem* 260, 5616 -20.

46. Murata, T., Bognar, A. L., Hayashi, T., Ohnishi, M., Nakayama, K. & Terawaki, Y. (2000). Molecular analysis of the folC gene of *Pseudomonas aeruginosa*. *Microbiol Immunol* 44, 879-86.
47. Post, D. A., Hove-Jensen, B. & Switzer, R. L. (1993). Characterization of the hemA-prs region of the *Escherichia coli* and *Salmonella typhimurium* chromosomes: identification of two open reading frames and implications for prs expression. *J Gen Microbiol* 139, 259-66.
48. Ravcheev, D. A., Gel'fand, M. S., Mironov, A. A. & Rakhmaninova, A. B. (2002). [Purine regulon of gamma-proteobacteria: a detailed description]. *Genetika* 38, 1203-14.
49. Reizer, J., Reizer, A. & Saier, M. H., Jr. (1995). Novel phosphotransferase system genes revealed by bacterial genome analysis--a gene cluster encoding a unique Enzyme I and the proteins of a fructose-like permease system. *Microbiology* 141 (Pt 4), 961-71.
50. Reizer, J., Ramseier, T. M., Reizer, A., Charbit, A. & Saier, M. H., Jr. (1996). Novel phosphotransferase genes revealed by bacterial genome sequencing: a gene cluster encoding a putative N-acetylgalactosamine metabolic pathway in *Escherichia coli*. *Microbiology* 142 (Pt 2), 231-50.
51. Soutourina, J., Blanquet, S. & Plateau, P. (2001). Role of D-cysteine desulphydrase in the adaptation of *Escherichia coli* to D-cysteine. *J Biol Chem* 276, 40864-72.
52. Troup, B., Jahn, M., Hungerer, C. & Jahn, D. (1994). Isolation of the hemF operon containing the gene for the *Escherichia coli* aerobic coproporphyrinogen III oxidase by in vivo complementation of a yeast HEM13 mutant. *J Bacteriol* 176, 673-80.
53. Wei, Y., Lee, J. M., Smulski, D. R. & LaRossa, R. A. (2001). Global impact of sdiA amplification revealed by comprehensive gene expression profiling of *Escherichia coli*. *J Bacteriol* 183, 2265-72.
54. Weng, M., Makaroff, C. A. & Zalkin, H. (1986). Nucleotide sequence of *Escherichia coli* pyrG encoding CTP synthetase. *J Biol Chem* 261, 5568-74.
55. Xie, G., Brettin, T. S., Bonner, C. A. & Jensen, R. A. (1999). Mixed-function supraoperons that exhibit overall conservation, albeit shuffled gene organization, across wide intergenomic distances within eubacteria. *Microb Comp Genomics* 4, 5-28.
56. Yew, W. S. & Gerlt, J. A. (2002). Utilization of L-ascorbate by *Escherichia coli* K-12: assignments of functions to products of the yjf sga and yia-sgb operons. *J Bacteriol* 184, 302-6.
57. Huynen, M. A. & Bork, P. (1998). Measuring genome evolution. *Proc Natl Acad Sci U S A* 95, 5849-56.
58. Lathe, W. C., 3rd, Snel, B. & Bork, P. (2000). Gene context conservation of a higher order than operons. *Trends Biochem Sci* 25, 474-9.
59. Goldschmidt, E. P., Cater, M. S., Matney, T. S., Butler, M. A. & Greene, A. (1970). Genetic analysis of the histidine operon in *Escherichia coli* K12. *Genetics* 66, 219-29.
60. Eisenstein, E., Gilliland, G. L., Herzberg, O., Moulton, J., Orban, J., Poljak, R. J., Banerjee, L., Richardson, D. & Howard, A. J. (2000). Biological function

- made crystal clear - annotation of hypothetical proteins via structural genomics. *Curr Opin Biotechnol* 11, 25-30.
61. Gilliland, G. L., Teplyakov, A., Obmolova, G., Tordova, M., Thanki, N., Ladner, J., Herzberg, O., Lim, K., Zhang, H., Huang, K., Li, Z., Tempczyk, A., Krajewski, W., Parsons, L., Yeh, D. C., Orban, J., Howard, A. J., Eisenstein, E., J. F. P., Bonander, N., Fisher, K. E., Toedt, J., Reddy, P., Rao, C. V., Melamud, E. & Moulton, J. (2002). Assisting functional assignment for hypothetical *Haemophilus influenzae* gene products through structural genomics. *Curr Drug Targets Infect Disord* 2, 339-53.
 62. Parsons, J. F., Lim, K., Tempczyk, A., Krajewski, W., Eisenstein, E. & Herzberg, O. (2002). From structure to function: YrbI from *Haemophilus influenzae* (HI1679) is a phosphatase. *Proteins* 46, 393-404.
 63. Tzeng, Y. L., Datta, A., Strole, C., Kolli, V. S., Birck, M. R., Taylor, W. P., Carlson, R. W., Woodard, R. W. & Stephens, D. S. (2002). KpsF is the arabinose-5-phosphate isomerase required for 3-deoxy-D-manno-octulosonic acid biosynthesis and for both lipooligosaccharide assembly and capsular polysaccharide expression in *Neisseria meningitidis*. *J Biol Chem* 277, 24103-13.
 64. Wu, J., Kasif, S. & DeLisi, C. (2003). Identification of functional links between genes using phylogenetic profiles. *Bioinformatics* 19, 1524-30.
 65. Lim, K., Tempczyk, A., Parsons, J. F., Bonander, N., Toedt, J., Kelman, Z., Howard, A., Eisenstein, E. & Herzberg, O. (2003). Crystal structure of YbaB from *Haemophilus influenzae* (HI0442), a protein of unknown function coexpressed with the recombinational DNA repair protein RecR. *Proteins* 50, 375-9.
 66. Voloshin, O. N., Ramirez, B. E., Bax, A. & Camerini-Otero, R. D. (2001). A model for the abrogation of the SOS response by an SOS protein: a negatively charged helix in DinI mimics DNA in its interaction with RecA. *Genes Dev* 15, 415-27.
 67. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* 96, 4285-8.
 68. McGuffin, L. J., Street, S. A., Bryson, K., Sorensen, S. A. & Jones, D. T. (2004). The Genomic Threading Database: a comprehensive resource for structural annotations of the genomes from key organisms. *Nucleic Acids Res* 32, D196-9.
 69. Bowers, P. M., Pellegrini, M., Thompson, M. J., Fierro, J., Yeates, T. O. & Eisenberg, D. (2004). Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol* 5, R35.
 70. Marcotte, C. J. & Marcotte, E. M. (2002). Predicting functional linkages from gene fusions with confidence. *Appl Bioinformatics* 1, 93-100.
 71. Wu, J. & Woodard, R. W. (2003). *Escherichia coli* YrbI is 3-deoxy-D-manno-octulosonate 8-phosphate phosphatase. *J Biol Chem* 278, 18117-23.
 72. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nature* 402, 83-6.

73. Berman, H. M. & Westbrook, J. D. (2004). The impact of structural genomics on the protein data bank. *Am J Pharmacogenomics* 4, 247- 52.
74. Browne, W. J., North, A. C., Phillips, D. C., Brew, K., Vanaman, T. C. & Hill, R. L. (1969). A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. *J Mol Biol* 42, 65-86.
75. Brenner, S. E. (2001). A tour of structural genomics. *Nat Rev Genet* 2, 801-9.
76. Wolf, Y. I., Grishin, N. V. & Koonin, E. V. (2000). Estimating the number of protein folds and families from complete genome data. *J Mol Biol* 299, 897-905.
77. Orengo, C. A., Todd, A. E. & Thornton, J. M. (1999). From protein structure to function. *Curr Opin Struct Biol* 9, 374-82.
78. Sternberg, M. J., Bates, P. A., Kelley, L. A. & MacCallum, R. M. (1999). Progress in protein structure prediction: assessment of CASP3. *Curr Opin Struct Biol* 9, 368-73.
79. Liu, J. & Rost, B. (2002). Target space for structural genomics revisited. *Bioinformatics* 18, 922-33.
80. Tramontano, A. & Morea, V. (2003). Assessment of homology-based predictions in CASP5. *Proteins* 53 Suppl 6, 352-68.
81. Baker, D. & Sali, A. (2001). Protein structure prediction and structural genomics. *Science* 294, 93-6.
82. Venclovas, C., Zemla, A., Fidelis, K. & Moult, J. (2003). Assessment of progress over the CASP experiments. *Proteins* 53 Suppl 6, 585-95.
83. Nagano, N., Orengo, C. A. & Thornton, J. M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J Mol Biol* 321, 741-65.
84. Dayhoff, M. O. (1976). The origin and evolution of protein superfamilies. *Fed Proc* 35, 2132-8.
85. Haft, D. H., Selengut, J. D. & White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Res* 31, 371-3.
86. Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D. & Kahn, D. (2002). ProDom: automated clustering of homologous domains. *Brief Bioinform* 3, 246-51.
87. Gracy, J. & Argos, P. (1998). Automated protein sequence database classification. I. Integration of compositional similarity search, local similarity search, and multiple sequence alignment. *Bioinformatics* 14, 164-73.
88. Gracy, J. & Argos, P. (1998). Automated protein sequence database classification. II. Delineation Of domain boundaries from sequence similarities. *Bioinformatics* 14, 174-87.
89. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30, 1575-84.
90. Mohseni-Zadeh, S., Brezellec, P. & Risler, J. L. (2004). Cluster-C, an algorithm for the large-scale clustering of protein sequences based on the extraction of maximal cliques. *Comput Biol Chem* 28, 211-8.
91. Gough, J. (2002). The SUPERFAMILY database in structural genomics. *Acta Crystallogr D Biol Crystallogr* 58, 1897-900.

92. Heger, A. & Holm, L. (2001). Picasso: generating a covering set of protein family profiles. *Bioinformatics* 17, 272-9.
93. Yona, G., Linial, N. & Linial, M. (2000). ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res* 28, 49-55.
94. Wu, C. H., Xiao, C., Hou, Z., Huang, H. & Barker, W. C. (2001). iProClass: an integrated, comprehensive and annotated protein classification database. *Nucleic Acids Res* 29, 52-4.
95. Ohlson, T., Wallner, B. & Elofsson, A. (2004). Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods. *Proteins* 57, 188-97.
96. Edgar, R. C. & Sjolander, K. (2004). COACH: profile-profile alignment of protein families using hidden Markov models. *Bioinformatics* 20, 1309-18.
97. Sadreyev, R. & Grishin, N. (2003). COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol* 326, 317-36.
98. Brenner, S. E., Chothia, C. & Hubbard, T. J. (1997). Population statistics of protein structures: lessons from structural classifications. *Curr Opin Struct Biol* 7, 369-76.
99. Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature* 357, 543-4.
100. Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science* 273, 595-603.
101. Wang, Z. X. (1998). A re-estimation for the total numbers of protein folds and superfamilies. *Protein Eng* 11, 621-6.
102. Zhang, C. & DeLisi, C. (1998). Estimating the number of protein folds. *J Mol Biol* 284, 1301-5.
103. Govindarajan, S., Recabarren, R. & Goldstein, R. A. (1999). Estimating the total number of protein folds. *Proteins* 35, 408-14.
104. Grant, A., Lee, D. & Orengo, C. (2004). Progress towards mapping the universe of protein folds. *Genome Biol* 5, 107.
105. Liu, J. & Rost, B. (2004). Sequence-based prediction of protein domains. *Nucleic Acids Res* 32, 3522-30.
106. Wootton, J. C. (1994). Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* 18, 269-85.
107. Liu, J. & Rost, B. (2004). CHOP: parsing proteins into structural domains. *Nucleic Acids Res* 32, W569-71.
108. Pipenbacher, P., Schliep, A., Schneckener, S., Schonhuth, A., Schomburg, D. & Schrader, R. (2002). ProClust: improved clustering of protein sequences with an extended graph-based approach. *Bioinformatics* 18 Suppl 2, S182-91.
109. Yona, G., Linial, N. & Linial, M. (1999). ProtoMap: automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins* 37, 360-78.
110. Karsay, R. Y., Wang, G., Dongre, N., Gao, G. & Dunbrack, R. L., Jr. (2002). CASA: a server for the critical assessment of protein sequence alignment accuracy. *Bioinformatics* 18, 496-7.

111. Brenner, S. E., Chothia, C. & Hubbard, T. J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci U S A* 95, 6073-8.
112. Park, J., Teichmann, S. A., Hubbard, T. & Chothia, C. (1997). Intermediate sequences increase the detection of homology between sequences. *J Mol Biol* 273, 349-54.
113. Park, J., Holm, L., Heger, A. & Chothia, C. (2000). RSDB: representative protein sequence databases have high information content. *Bioinformatics* 16, 458-64.
114. Enright, A. J. & Ouzounis, C. A. (2000). GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics* 16, 451-7.
115. Karplus, K. & Hu, B. (2001). Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. *Bioinformatics* 17, 713-20.
116. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 305, 567-80.
117. Marchler-Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A., Geer, L. Y. & Bryant, S. H. (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res* 30, 281-3.
118. Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol* 284, 1201-10.
119. Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F. & Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 29, 291-325.
120. Liu, J., Hegyi, H., Acton, T. B., Montelione, G. T. & Rost, B. (2004). Automatic target selection for structural genomics on eukaryotes. *Proteins* 56, 188-200.
121. Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *Embo J* 5, 823-6.
122. Coulson, A. F. & Moulton, J. (2002). A unfold, mesofold, and superfold model of protein fold use. *Proteins* 46, 61-71.
123. Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C. & Teichmann, S. A. (2004). Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol* 14, 208-16.
124. Bashton, M. & Chothia, C. (2002). The geometry of domain combination in proteins. *J Mol Biol* 315, 927-39.
125. DeWeese-Scott, C. & Moulton, J. (2004). Molecular modeling of protein function regions. *Proteins* 55, 942-61.
126. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris,

- W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Showkneen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.
127. Lawrence, J. G. & Ochman, H. (1998). Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci U S A* 95, 9413-7.
128. Goodman, S. D. & Scocca, J. J. (1988). Identification and arrangement of the DNA sequence recognized in specific transformation of *Neisseria gonorrhoeae*. *Proc Natl Acad Sci U S A* 85, 6982-6.
129. Elkins, C., Thomas, C. E., Seifert, H. S. & Sparling, P. F. (1991). Species-specific uptake of DNA by gonococci is mediated by a 10-base-pair sequence. *J Bacteriol* 173, 3911-3.
130. Smith, H. O., Tomb, J. F., Dougherty, B. A., Fleischmann, R. D. & Venter, J. C. (1995). Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome. *Science* 269, 538-40.
131. Cheetham, B. F. & Katz, M. E. (1995). A role for bacteriophages in the evolution and transfer of bacterial virulence determinants. *Mol Microbiol* 18, 201-8.
132. Groisman, E. A. & Ochman, H. (1996). Pathogenicity islands: bacterial evolution in quantum leaps. *Cell* 87, 791-4.
133. Hacker, J. & Kaper, J. B. (2000). Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol* 54, 641-79.
134. Ochman, H., Lawrence, J. G. & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299-304.
135. Brown, N. L. & Evans, L. R. (1991). Transposition in prokaryotes: transposon Tn501. *Res Microbiol* 142, 689-700.
136. Scott, J. R. & Churchward, G. G. (1995). Conjugative transposition. *Annu Rev Microbiol* 49, 367-97.
137. Kuduvali, P. N., Rao, J. E. & Craig, N. L. (2001). Target DNA structure plays a critical role in Tn7 transposition. *Embo J* 20, 924-32.
138. Eisen, J. A. (2000). Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Curr Opin Genet Dev* 10, 606-11.
139. Lawrence, J. G. & Ochman, H. (2002). Reconciling the many faces of lateral gene transfer. *Trends Microbiol* 10, 1-4.

140. Medigue, C., Rouxel, T., Vigier, P., Henaut, A. & Danchin, A. (1991). Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J Mol Biol* 222, 851-6.
141. Lawrence, J. G. & Ochman, H. (1997). Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* 44, 383-97.
142. Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., McDonald, L., Utterback, T. R., Malek, J. A., Linher, K. D., Garrett, M. M., Stewart, A. M., Cotton, M. D., Pratt, M. S., Phillips, C. A., Richardson, D., Heidelberg, J., Sutton, G. G., Fleischmann, R. D., Eisen, J. A., Fraser, C. M. & et al. (1999). Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* 399, 323-9.
143. Worning, P., Jensen, L. J., Nelson, K. E., Brunak, S. & Ussery, D. W. (2000). Structural analysis of DNA sequence: evidence for lateral gene transfer in *Thermotoga maritima*. *Nucleic Acids Res* 28, 706-9.
144. Koonin, E. V., Makarova, K. S. & Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol* 55, 709-42.
145. Salzberg, S. L., White, O., Peterson, J. & Eisen, J. A. (2001). Microbial genes in the human genome: lateral transfer or gene loss? *Science* 292, 1903-6.
146. Dobrindt, U., Hochhut, B., Hentschel, U. & Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol* 2, 414-424.
147. Fitzgerald, J. R. & Musser, J. M. (2001). Evolutionary genomics of pathogenic bacteria. *Trends Microbiol* 9, 547-53.
148. Andersson, J. O. & Roger, A. J. (2003). Evolution of glutamate dehydrogenase genes: evidence for lateral gene transfer within and between prokaryotes and eukaryotes. *BMC Evol Biol* 3, 14.
149. Grantham, R., Gautier, C., Gouy, M., Mercier, R. & Pave, A. (1980). Codon catalog usage and the genome hypothesis. *Nucleic Acids Res* 8, r49-r62.
150. Groisman, E. A., Saier, M. H., Jr. & Ochman, H. (1992). Horizontal transfer of a phosphatase gene as evidence for mosaic structure of the *Salmonella* genome. *Embo J* 11, 1309-16.
151. Garcia-Vallve, S., Romeu, A. & Palau, J. (2000). Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res* 10, 1719-25.
152. Sharp, P. M. & Li, W. H. (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15, 1281-95.
153. Koski, L. B., Morton, R. A. & Golding, G. B. (2001). Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol Biol Evol* 18, 404-12.
154. Olendzenski, L., Liu, L., Zhaxybayeva, O., Murphey, R., Shin, D. G. & Gogarten, J. P. (2000). Horizontal transfer of archaeal genes into the deinococcaceae: detection by molecular and computer-based approaches. *J Mol Evol* 51, 587-99.

155. Coombs, J. M. & Barkay, T. (2004). Molecular evidence for the evolution of metal homeostasis genes by lateral gene transfer in bacteria from the deep terrestrial subsurface. *Appl Environ Microbiol* 70, 1698-707.
156. Fitch, W. M. & Farris, J. S. (1974). Evolutionary trees with minimum nucleotide replacements from amino acid sequences. *J Mol Evol* 3, 263-78.
157. Dallal, G. E. & Rousseeuw, P. J. (1992). LMSMVE: a program for least median of squares regression and robust distances. *Comput Biomed Res* 25, 384-91.
158. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22, 4673-80.
159. Mirkin, B. G., Fenner, T. I., Galperin, M. Y. & Koonin, E. V. (2003). Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol* 3, 2.
160. Balch, W. E., Magrum, L. J., Fox, G. E., Wolfe, R. S. & Woese, C. R. (1977). An ancient divergence among the bacteria. *J Mol Evol* 9, 305-11.
161. Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E. & Stanhope, M. J. (2001). Universal trees based on large combined protein sequence data sets. *Nat Genet* 28, 281-5.
162. Dutilh, B. E., Huynen, M. A., Bruno, W. J. & Snel, B. (2004). The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. *J Mol Evol* 58, 527-39.
163. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8, 275-82.
164. Johansen, B. K., Wasteson, Y., Granum, P. E. & Brynstad, S. (2001). Mosaic structure of Shiga-toxin-2-encoding phages isolated from Escherichia coli O157:H7 indicates frequent gene exchange between lambdoid phage genomes. *Microbiology* 147, 1929-36.
165. Ragan, M. A. (2001). On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol Lett* 201, 187-91.