UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Evolution studied through protein structural domains

Permalink

https://escholarship.org/uc/item/0zm8m8zw

Author

Yang, Song

Publication Date

2007

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Evolution Studied Through Protein Structural Domains

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in

Chemistry

by

Song Yang

Committee in charge:

Professor Philip E. Bourne, Chair
Professor Douglas Magde, Co-Chair
Professor Russell F. Doolittle
Professor Simpson Joseph
Professor Shankar Subramaniam

2007

Copyrights©

Song Yang, 2007

All rights reserved.

The dissertation of Song Yang is approved, and it is acceptable in quality and form for publication on microfilm.

_____

_____

_____

_____

Co-Chair

_____

Chair

University of California, San Diego

2007

TABLE OF CONTENTS

Signature Page……iii

Table of Contents……iv

List of Figures……....vii

List of Tables……...……x

Acknowledgement……..xi

Vita……xii

Abstract ……..……..xiv

Chapter 1 Introduction 1

1.1 Protein Sequence, Structure and the Protein Universe 1
1.1.1 Protein Sequence and the Protein Universe 1
1.1.2 Protein Structure and Sequence-Structure Relationship 3
1.2 Protein Domain 8
1.2.1 The Definition of Protein Domain 8
1.2.2 The Properties of Protein Domains 11
1.2.3 Protein Domain Classification 13
1.3 Protein Domain as Evolutionary Unit 16
1.3.1 The Power Law Distribution of Protein Domains 16
1.3.2 Protein Domain Analysis in the Era of Complete Genomes 17
1.3.3 The History of Protein Domains 19
1.4 References 21

Chapter 2 Data and Algorithms 26

2.1 Databases and Algorithms 26
2.1.1 PDB Data 26
2.1.2 SCOP Protein Structure Classification 26
2.1.3 Complete Genome Sequences 29
2.1.4 Hidden Markov Model 30
2.1.5 SUPERFAMILY Database 32
2.2 Lists of Organisms in Study 35
2.2.1 SUPERFAMILY Release 2004-07-25 35
2.2.2 SUPERFAMILY Release 2005-10-09 43
2.3 References 47

Chapter 3 Phylogeny by Protein Domain Content 49

3.1 Introduction 49
3.2 Method 51
3.2.1 Data Sources 51
3.2.2 Data Management 53
3.2.3 Phylogenetic Methods 55
3.3 Phylogenetic Trees 56
3.3.1 Results 56
3.3.2 Discussion 69
3.4 Distribution of Protein Domains in the Three Superkingdoms 75
3.4.1 Results 75
3.4.2 Discussion 85
3.5 Conclusion 90
3.6 References 91

Chapter 4 Genome Alignment Using Protein Domain 94

4.1 Introduction 94
4.2 Method 96

4.2.1 Data Management 96
4.2.2 Domain Dot Plot 96
4.2.3 Dynamic Programming 97
4.3 Results 98
4.3.1 Genome Evolution Viewed by Domain Dot Plots 98
4.3.2 Symmetrical inversion of prokaryote genomes around the replication origin 102
4.4 Discussion 106
4.4.1 Domains as Evolutionary Units 106
4.4.2 Future Directions 107
4.5 References 108

Chapter 5 Evolution of Protein Domains 109

5.1 Introduction 109
5.1.1 The Evolutionary History of Protein Domains 109
5.1.2 Previous Works 111
5.1.3 The Hypothesis to Be Tested 113
5.2 Domain Tree and Domain Combination Tree 115
5.2.1 Methods 115
5.2.2 Results 125
5.2.3 Discussion 138
5.3 Systematically Derive the Origin and Evolution of Domains 144
5.3.1 Method 145
5.3.2 Results 149
5.3.3 Conclusion 157
5.4 References 158

LIST OF FIGURES

Figure 1.1 Comparison of the growth rate of sequences in Genbank and structures in PDB in last two decades. Figures were obtained from the Genbank (http://www.ncbi.nlm.nih.gov/ Genbank/) and PDB (http://www.rcsb.org/pdb/) websites. 4

Figure 1.2 Divergence of two sequences from a common ancestor and methods of detection. 7

Figure 1.3 (a) Domain architecture and (b) Tertiary structure of hematopoetic cell kinase (PDB code 1qcf), containing protein kinase and src homology 2 and 3 (SH2 and SH3) domains [19]. 9

Figure 1.4 Representations of the domain architectures of human p60 tim (transforming immortalized mammary oncogene), mouse βPix-c, human Vav oncogene, and human RGS1 (regulator of G-protein signalling 1). Domain abbreviations: C1, protein kinase C conserved region 1; C2, protein kinase C conserved region 2; CH, calponin homology; PH, pleckstrin homology; RasGAP, GTPase activator protein specific for Ras-like small GTPases; RhoGEF, guanine nucleotide exchange factor specific for Rho-type small GTPases; SH2, src homology 2; SH3, src homology 3. Species abbreviations: Hs, Homo sapiens; Mm, Mus musculus. [19] 11

Figure 1.5 Double-logarithmic plot of the distribution of protein folds by the number of families [3]. 16

Figure 2.1 A linear hidden Markov model and example alignment 31

Figure 3.1 The comparison of Fold number vs. e-value cutoff figure between different release of the SUPERFAMILY database. (A) The genomic domain assignment of early version of SUPERFAMILY (Jan 06, 2004). (B) The genomic domain assignment of current version of SUPERFAMILY (July 25, 2004) 52

Figure 3.2 The illustration of the protein domains shared by closely related free-living and parasitic organisms. 54

Figure 3.3 NJ tree of 36 Eukaryote based exclusively on presence or absence of SCOP superfamily folds. Bootstrap numbers are limited to the major nodes. 57

Figure 3.4 The unweighted neighbor-joining tree with the following types of protein domains removed from X. tropicalis: (A) 19 non-Eukaryote domains; (B) 40 non-animal domains; (C) 74 non-vertebrate domains. 58

Figure 3.5 NJ tree of 19 Archaea based exclusively on presence or absence of SCOP superfamily folds. Bootstrap numbers are limited to the major nodes. 61

Figure 3.6 NJ tree (with weighting) of 119 Bacteria based exclusively on the presence or absence of SCOP superfamily folds. Asterisks denote anomalously positioned taxa. 63

Figure 3.7 Unweighted (A) and weighted (B) PDC-based phylogenetic tree of 123 Bacteria organisms. All the major lineages at the order level are denoted by different color. The genome sizes of all species are listed. 64

Figure 3.8 Overall phylogeny (NJ) of 174 organisms for which complete genomes have been determined. Bootstrap number was limited to the major branch point. 68

Figure 3.9 The NJ (A) and UPGMA (B) Eukaryote trees based on protein domain abundance. Plant species that have large genome size and thus more protein domains protrude from the rest of the organism in the tree. 71

Figure 3.10 Venn diagrams showing occurrence of 1,244 Domain Superfamilies in the three superkingdoms. The first value in any sector reflects the occurrence of a particular Domain Superfamilies in any genome in that sector; values in parentheses indicate numbers found in all genomes in a sector. (A) Census based on 174 complete genomes. (B) Census based on complete genomes of 19 Archaea, 19 selected Eukarya, and 19 selected Bacteria. 75

Figure 3.11 Structural comparison of the six Superfamilies in the d.17 Fold 81

Figure 3.12 Numbers and percentages of disulfide bond-dependent domains in each kingdom. The criteria for disulfide bond-dependent is over 70% of the PDB representatives contain disulfide bond. 84

Figure 3.13 Simulation of the decay curve of the number of unique and ubiquitous protein domain superfamilies when all 174 genomes are randomly sampled without repeat 1000 times. 88

Figure 4.1 (A) Domain dot plot of two *Salmonella enterica* strains, *CT18* (vertical) and *Ty2* (Horizontal). Blue dots represent domain matches, red dots represent the calculated domain local alignment and green dots represent extrapolated local alignment. (B) The chromosome inversion determined by comparative genome analysis [7]. 99

Figure 4.2 Domain dot plot for Chlamydophila pneumoniae AR39 and CWL029 represents a major chromosome translocation. 100

Figure 4.3 Comparison of the X chromosomes of human and mouse by (A) Protein domain comparison. (B) Genome comparison. Boxes in the plot represent proteins with tandem repeat domains, which exist extensively in higher eukaryotes [8]. ... 101

Figure 4.4 The genome alignment of three pairs of closely related species, each pair involves various genome inversion. 103

Figure 4.5 Multiple chromosome inversions arround the replication origin leads to the X-shape genome alignment observed in the dot plot. Replication origins of the circular chromosome are located at the start position of the plot. Arrows at the top of each plot mark the location of inversion. 105

Figure 5.1 The Venn diagram of the distribution of folds in the three superkingdoms of life (left) and the corresponding phylogenetic tree at the root of the life (right). ... 114

Figure 5.2 (A) a schematic phylogenetic tree represented by (B) a text file in the standard tree format. The tree is generated by the TreeView software. 116

Figure 5.3 Protein domain length distribution according to SCOP 1.69. Domain definitions are retrieved from the ASTRAL database with 95% sequence identity. 120

Figure 5.4 Sequence comparison of two lactate dehydrogenases using the Blastp program confirms that they are homologous proteins. The analysis uses the NCBI Blast 2 sequence server, which enables the pair wise sequence comparison by Blast. 122

Figure 5.5 The domain trees of the pilin family (d.42.1.1) and the TcpA-like family (d.42.1.2). Both families exist in bacteria exclusively, and only the proteobacteria taxa within the bacteria are shown due to lack of space. The number behind each species represents the abundance of the domain family in the genome of the species. The d.42.1.1 family exists in most of the proteobacteria species, except in several α proteobacteria and γ proteobacteria clades that might result from gene loss. The d.42.1.2 family only exists in Vcho and Vfis, which are marked in the figure. 127

Figure 5.6 The domain combination Eukaryotic tree of a.109.1.1, the Class II MHC-associated invariant chain ectoplasmic trimerization domain. The letters next to the species represent different combination types. In this case, type b corresponds to N/A ~ a.109.1.1 and c represents N/A ~ a.109.1.1 ~ g.28.1.1, where N/A is an unknown domain. 130

Figure 5.7 Comparison of the 3D structures of the remotely related domain combinations that contain the same domain architecture at the superfamily level but differ at the family level. The b.1.10 domains on the left contain two parallel beta-sheets, and the d.105.1containing one beta sheet locates on the right. 134

Figure 5.8 The domain combination Eukaryotic trees of three clathrin adaptor appendage domain families. 135

Figure 5.9 The PDB-validated archaeal trees of two domain combination, d.58.31.2 ~ a.89.1.1 and d.58.31.1, are compared. The numbers and types in the parenthesis, such as '<Mkan 2b>', represent the predicted domain combinations; those outside of the parenthesis, such as 'Methanopyrus kandleri 2b', represent the PDB containing combinations that have solved 3D structures from that species. 136

Figure 5.10 The PDB-validated domain combination trees of phycocyanin-like phycobilisome proteins (a.1.1.3). (A) Part of the bacteria tree, a.1.1.3 exists only in cyanobactria species. 141

Figure 5.11 Two possible evolutionary histories for domains that result in the same current domain distribution 147

Figure 5.12 The four possible evolutionary events between a pair of parents/children. 148

Figure 5.13 The histogram and the double-log plot of domain combination distribution in species with complete genomes shows a power-law distribution. 150

Figure 5.14 The histogram and the double-log plot of the number of combinations that each domain is involve in. 152

Figure 5.15 The predicted number of domains and domain combinations that originated at each node on the eukaryotic tree. 156

LIST OF TABLES

Table 2.1 SCOP Classification Statistics 29

Table 2.2 The SCOP version and number of annotated complete genomes in each release of the SUPERFAMILY database. Some complete genomes belong to different strains of the same species. The number of unique genomes in each category only considers a single strain of any given species. 33

Table 2.3 The complete genomes in SUPERFAMILY database release 2004-07-25. 35

Table 2.4 The genomes added in the SUPERFAMILY since release 2004-07-25 43

Table 3.1 (A) Data matrix of presence/absence of protein domains in various species ... 53

Table 3.2 (A) Unique Superfamilies exist exclusively within each region: A (1), E (14), AE (2), ABE (49) 77

Table 4.1 Comparison of evolution of protein sequence and chromosome architecture. 94

Table 5.1 Three examples of domain combinations. The boundary and identity of each domain are listed. 118

Table 5.2 Comparison of domain combination assignment for two homologous proteins (D-lactate dehydrogenase) in *Bfra2* and *Hpyl1*. 121

Table 5.3 The domain combinations of three clathrin adaptor appendage domain and their evolutionary origin in the tree of life. 132

Table 5.4 Penalty scores for evolutionary processes. 145

Table 5.5 The total types of domain combinations. The second and third rows list the number of domain combination before and after correction for errors. The fourth row lists combinations with high abundance. Specifically, for a certain domain, only the combinations that cover the top 95% of domain abundance are counted. 149

Table 5.6 The top 20 most abundant domain combinations in all. 151

Table 5.7 The top 10 domains with most number of combination types. 153

Table 5.8 The number of domain, domain combinations that originated from each nodes. The last column listed the ration of domain / combination. 155

ACKNOWLEDGEMENT

I would like to acknowledge Professor Phil Bourne for his support and guidance as my advisor through these years at UCSD, and Professor Russell Doolittle for his advice and help in my research. Without them, this work could not be accomplished.

Chapter 3 contains the material as it appears in Yang, S., R.F. Doolittle, and P.E. Bourne, Phylogeny determined by protein domain content. Proc Natl Acad Sci U S A, 2005. 102(2): p. 373-8. The dissertation author was the primary investigator and author of this paper.

VITA

1997	Bachelor of Science, Chemistry, Beijing University
2000	Master of Science, Physical Chemistry, Dalian Institute of Chemical Physics, Chinese Academy of Sciences
2000 – 2007	Research Assistant, University of California, San Diego
2007	Doctor of Philosophy, Chemistry, University of California, San Diego

PUBLICATIONS

1. Yang, S., R.F. Doolittle, and P.E. Bourne, Phylogeny determined by protein domain content. Proc Natl Acad Sci U S A, 2005. 102(2): p. 373-8.
2. Dupont, C.L., et al., Modern proteomes contain putative imprints of ancient shifts in trace metal geochemistry. Proc Natl Acad Sci U S A, 2006. 103(47): p. 17822-7.
3. Song Yang, Phil Bourne, Evolution of protein domain viewed by domain tree and domain combination tree. In preparation.

CONFERENCE

1. Song Yang, Russell Doolittle, Phil E. Bourne. The study of Evolution through Protein Domain Structure. ISMB 2004 3Dsig invited speaker.

POSTER

1. Song Yang, Phil Bourne. When and Where Have Protein Folds Emerged? An Evolutionary View. RECOMB 2004
2. Song Yang, Phil Bourne. Distribution of Protein Folds in the Three Major Superkingdoms of Life. ISMB 2004

ABSTRACT OF THE DISSERTATION

Evolution Studied Through Protein Structural Domains

by

Song Yang

Doctor of Philosophy in Chemistry

University of California, San Diego, 2007

Professor Philip E. Bourne, Chair

Professor Douglas Magde, Co-chair

A protein structural domain is defined as a compact, spatially distinct part of a protein that can fold independently of neighboring sequences. Since the number of protein domains is limited, and protein domains are evolutionarily more conserved than protein sequences, protein domains play an important role in our understanding of the structure, function and evolution of proteins. As fundamental evolutionary units, protein

domains are associated with a variety of evolutionary processes such as domain genesis, loss, horizontal gene transfer (HGT) and domain combination. In the era of complete genomes, the number and types of protein domains in over 300 organisms across all major lineages of life can be retrieved using sophisticated domain assignment algorithms. Protein domain content is used as a structural attribute to determine the relatedness of organisms and reconstruct a phylogenetic tree of life that is comparable to the phylogeny by sequence analysis and combinations of gene content and gene order. The positions and sequence of protein domains along the chromosomes of closely-related species, as viewed by genome alignment plots, reflects genome rearrangement events such as duplication and inversion. The evolutionary history of individual domains and domain combinations can be illustrated by domain trees and combination trees. The evolutionary origins of all protein domains suggest that a large proportion of protein domains were invented at the root of life, and during evolution biological organisms tend to create new proteins and functions through recombination and duplication of existing domains, rather than creation of new domains *de novo*. Taken together this work shows the power of using protein structural domains to study evolution.

Chapter 1 Introduction

1.1 Protein Sequence, Structure and the Protein Universe

It is generally accepted that protein sequence determines protein structure and protein structure determines protein function.

1.1.1 Protein Sequence and the Protein Universe

According to the central dogma of molecular biology, the primary sequences of proteins, the major functional materials of the cell, are encoded by the genetic material DNA contained in the chromosomes and are translated via RNA. Comprising any combination of twenty amino acids, the total numbers of possible protein sequences, or 'protein universe', is practically infinite [1-4]. Assuming an average protein length of 200 amino acids, there can theoretically be 20^{200} different protein sequences, a number that exceeds the number of atoms in the Universe.

The theoretical protein sequence space is an overestimation of the protein universe, since not all possible polypeptide chains possess the proper physical and chemical properties to fold into structural proteins, and not all the proteins with 3D structure have the appropriate function to be selected during evolution. The narrower protein universe can be estimated by the encoded genes in all species that ever existed on earth. Assuming that there are 10 million species on earth and the average number of

genes in a complete genome is about 5000 (an intermediate number between prokaryote and eukaryote), there are totally $5x10^{10}$ unique protein sequences, which is still a substantially large number.

Moreover, the population of the protein universe is not randomly distributed, as can be evidenced by the existence of homologous proteins between different species and within the same species. The concept of a protein family, which was introduced in the 1960s by Margaret Dayhoff's group, denotes a group of protein sequences that share high sequence similarity and are likely to be evolutionarily related [5, 6]. During the course of evolution, the sequence of a protein gradually diverges because of random mutation and natural selection. If the homologous proteins in different species share the same function, they are called orthologs; proteins in the same species that evolve through divergence after the duplication of the ancestor protein are called paralogs. Protein families not only organize a vast number of sequentially and functionally related proteins into homologous groups, but also reflect their evolutionary relationship; therefore, they play a key role in phylogenetic analysis and functional annotation.

The homologous proteins in each protein family share similar sequences, and differ significantly from the proteins of other families. Thus, protein families form isolated patches in protein sequence space, with the majority of the sequence space being empty. This discrete distribution of sequence space comes from the fact that a functional protein must fold into a stable three dimensional structure under certain biological conditions, and the stable structure requires certain physical and chemical properties of the polypeptide chain that are determined by the amino acids sequence of the protein [7].

1.1.2 Protein Structure and Sequence-Structure Relationship

Despite the unfolded proteins, most functional proteins have a unique and stable 3-D structure. Although in principle, protein structure can be determined through *ab initio* theoretical calculation based on the protein's primary sequence, in practice, the so called protein folding problem has proved to be extremely hard to solve [7]. Currently our knowledge of protein structures comes mainly from experimental methods, especially X-ray crystallography and Nuclear Magnetic Resonance (NMR) and more recently cryoelectron microscopy. After decades of development, these techniques are so powerful that even giant molecules such as the ribosomal RNA large subunit ($> 10^6$ Dalton) [8] and protein complexes such as the photosynthesis reaction center II complex that comprises 18 protein components [9] can be solved. All the available protein structures are deposited in to the Protein Data Dank (PDB) [10], which contained 40479 structures at the end of 2006, with the addition of about 150 per week. Figure 1.1 shows two decades of growth in the number of protein structures in the PDB compared to the number of protein sequences (genes) in Genbank.

While in recent years the growth rate of PDB structures has been almost exponential, the growth in the number of sequences remains much greater than that of structures.

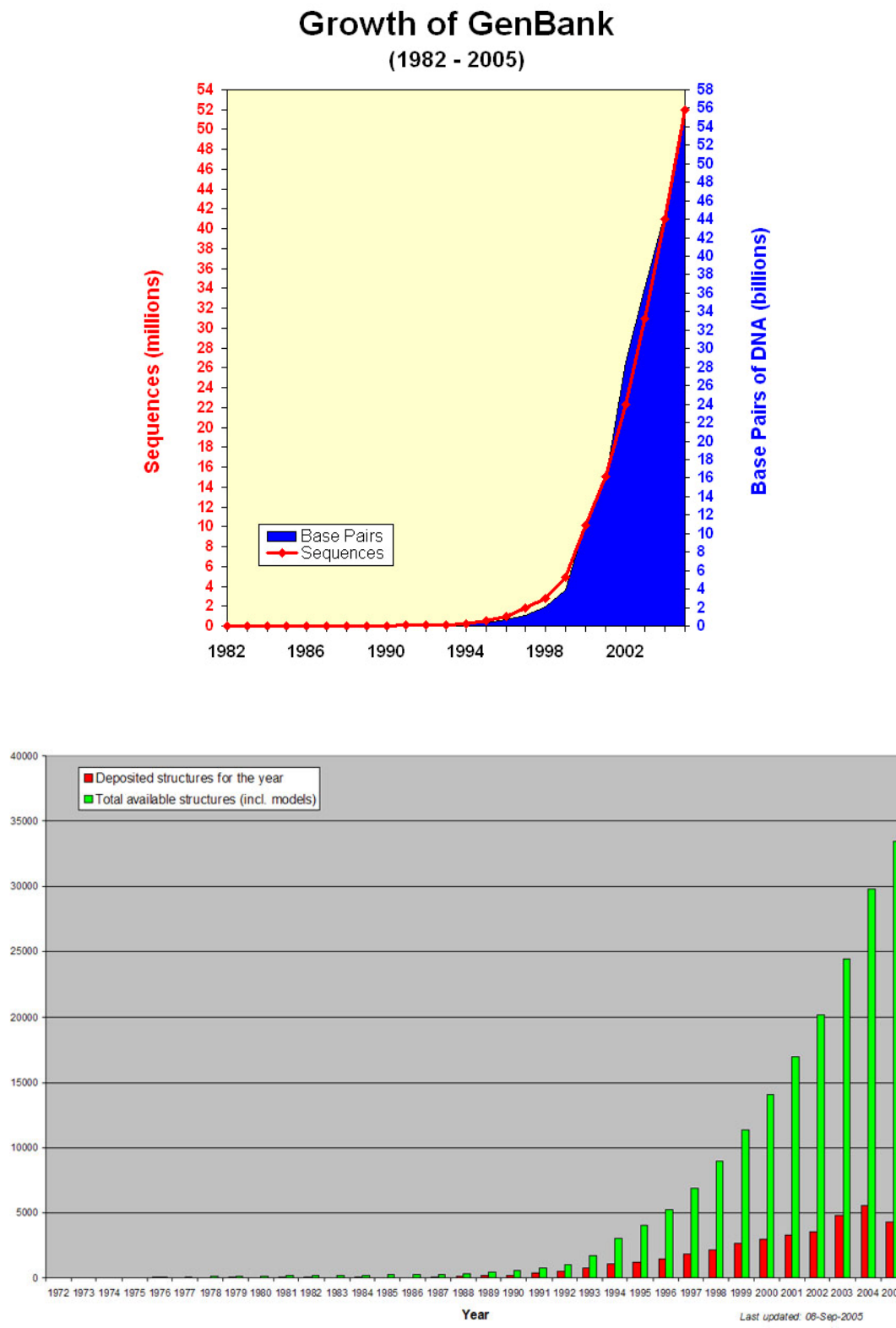


Figure 1.1 Comparison of the growth rate of sequences in Genbank and structures in PDB in last two decades. Figures were obtained from the Genbank (http://www.ncbi.nlm.nih.gov/ Genbank/) and PDB (http://www.rcsb.org/pdb/) websites.

Helped by high-throughput sequencing technologies, whole genome sequencing has become a mature and routine procedure, and is expected to be much cheaper and faster in the near future. Conversely, both X-ray crystallography and NMR have weaknesses in facilitating high-throughput structure determination of a genome-wide scale, structural genomics notwithstanding. NMR is applicable to relatively small proteins with molecular weight less than 50,000 Daltons, and the crystallization step is time-consuming and unpredictable in X-ray crystallography because some proteins such as the membrane proteins are difficult to crystallize [7].

Lack of a fast, universal and high-throughput structure determinant method means proteins with available 3-D structure represent only a tiny fraction of all known proteins. Moreover, among the over 40000 solved structures, many proteins are sequence homologs and share similar structures. For example, myoglobin, the first protein with a solved 3-D structure, is represented by over 300 entries in the PDB, with each position having been mutated. In short, the PDB contains a highly redundant set of structures.

Surprisingly, it was found that proteins with no apparent sequence similarity and sometimes with no apparent functional relationship have almost identical 3D structures. For instance, consider actin, a component of the cytoskeleton known only in eukaryotic cells. When its crystal structure was determined, it was found to resemble the ATPase domain of heat shock 70, a ubiquitous chaperon protein that exists in prokaryotes as well as in eukaryotes [11]. With only 16% sequence identity in 241 amino acid residues, the structurally equivalent parts of the two proteins can be superimposed with a root mean square distance (r.m.s.d) of 2.3 Å. Further comparison of the structure and function of

conserved key residues and motifs confirms that actin and the ATPase domain of heat shock 70 have a common ancestor and thus belong to the same family. Many other similar cases have been found, e.g., phycocyanin and globins [12], and cold shock protein and oligonucleotide-binding proteins [13]. These findings indicate that the structure of protein is more conserved throughout evolution than protein sequence and therefore allows us to recognize more distant ancestral relationships. It is this observation that is exploited throughout this work.

During evolution, proteins derived from a common ancestral protein can change their sequences and diverge by random mutations (including substitutions, insertions and deletions of residues), giving rise to families of homologous proteins. Not all positions are equally susceptible to mutation; some positions may be very important for function, stability, or folding, and others may be less constrained and diverge rapidly. If protein sequences diverge beyond the limit of the sequence similarity detection methods, the evolutionary relationship can not be elucidated by sequence signals alone. In this case, because of the restricted physical and chemical properties of proteins that facilitate their stable folding and function, those structures can still be conserved. This region is called the twilight zone [14] where the sequence similarity is blurred and the evolutionary relationship can only be recovered by structural information, as illustrated in Figure 1.2.

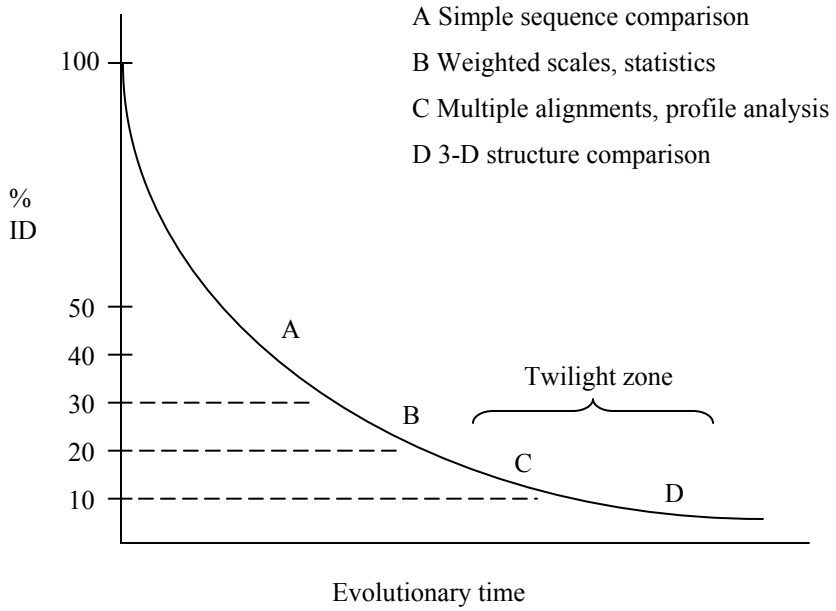


Figure 1.2 Divergence of two sequences from a common ancestor and methods of detection.

The protein family describes a group of homologous proteins with high sequence identity (>30%) and closely related in evolution. The concept of protein superfamily was introduced to define evolutionarily remotely related proteins that have little sequence similarity yet have a common ancestor, i.e., proteins in the twilight zone [5]. Although there is no clear boundary distinguishing families and superfamilies, a 30% sequence identity cutoff, which is close to the limit of the sequence comparison methods, has been widely accepted.

To complicate matters, the physical and chemical properties inherent in protein folding means that structural units (folds) can exist in proteins as a result of convergent evolution and hence have no evolutionary relationship but simply reflect the modularity inherent in protein structure.

1.2 Protein Domain

1.2.1 The Definition of Protein Domain

It has been known for a long time that some proteins contain spatially distinct regions [15]. In the 1960s, two of the first enzyme structures determined, lysozyme [16] and ribonuclease [17], were found to contain two distinct structural units, which were termed domains. Subsequently, more proteins were found to have multiple structural units, and it turned out that such domains could occur either in different structural contexts (such as the Rossmann folds in lactate dehydrogenase and alcohol dehydrogenase [18]) or in multiple copies in the same polypeptide chain (such as the fibronectin type I, II, III domain in fibronectin and other proteins). Figure 1.3 shows a multidomain protein, hematopoetic cell kinase that contains three structural units.

In defining protein structure, a domain is considered as a compact, spatially distinct part of a protein that can fold independently of neighboring sequences [7]. However, the meaning of domain varies in different biological contexts. In biochemistry, domains are often described as structural regions with distinct biological functions, irrespective of their three-dimensional structure. In sequence comparison, domains are used to describe sequence motifs or signature sequences. In many cases, these three views are compatible when the sequence homologs adopt similar folds and possess related functions. Fold is used to describe a unique tertiary structure that contains a certain topology of secondary structure elements. As fold is a characteristic of a protein domain, so sometimes fold and domain are used interchangeably.

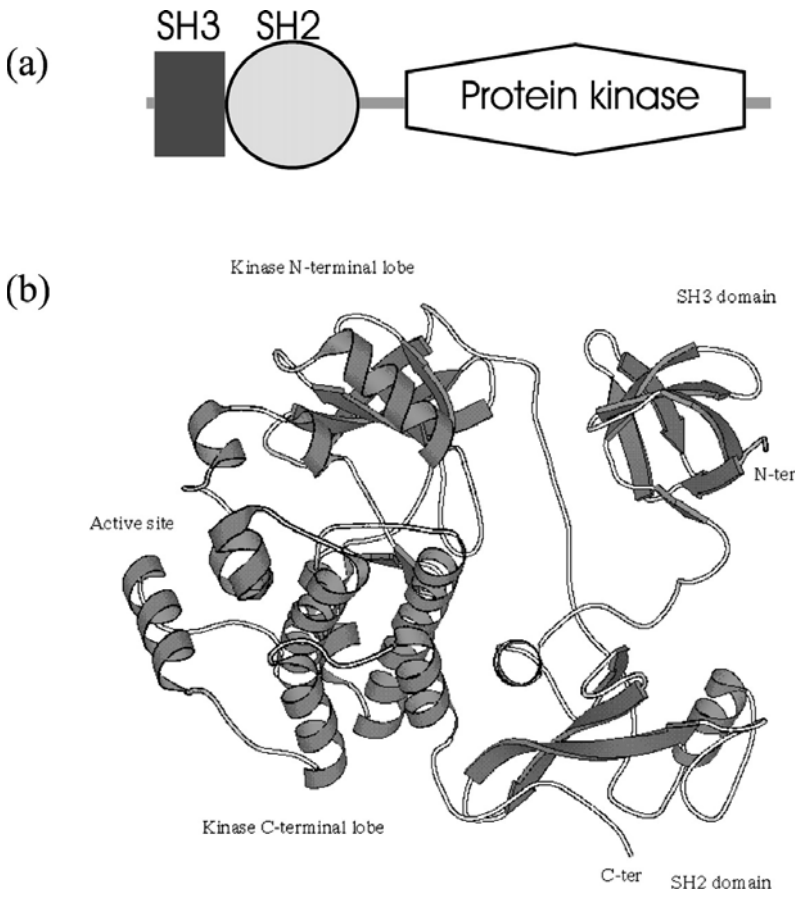


Figure 1.3 (a) Domain architecture and (b) Tertiary structure of hematopoetic cell kinase (PDB code 1qcf), containing protein kinase and src homology 2 and 3 (SH2 and SH3) domains [19].

Many methods assign protein domains according to a protein's 3D structure, most of which are based on the geometric measurement of compactness [20-23]. Helped by structure comparison methods, the observation of a similar structure in different structure contexts indicates the existence of a protein domain. The principle of recurrence plays a more important role in domain assignment methods that are based on sequence similarity, irrespective of whether or not a unit is known to be spatially distinct.

Different domain arrangements in protein sequences cause significant problems in sequence analysis. For example, the domain architecture of a group of pleckstrin homology (PH) and src homology 3 (SH3) containing proteins are listed in Figure 1.4. Among these proteins, the orientation of PH and SH3 can be either PH followed by SH3 or SH3 followed by PH, as well as the combination of SH2 and SH3: SH2 – SH3 – SH2 or SH3 – SH2 – SH3. When comparing the sequence of any one of these sequences with a sequence database, a large number of hits with a variety of domain architectures will be returned. Some hits match the entire domain arrangement of the protein; some only match parts of the multidomain protein; others only share one domain, and yet others are missed.

Such a set of conserved regions represents a family of homologous domains. Domain families represent a group of domains that have high sequence similarity and derive from a common ancestor. The concept of domain family extends the homologous relationship at the protein level into the sub-protein or domain level. Beyond domain assignment methods based on structure compactness, multiple protein sequence comparison is a popular way to define structural domains.

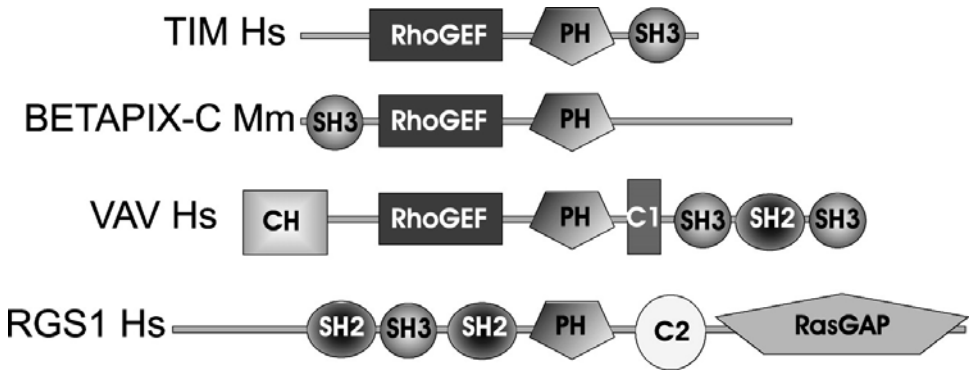


Figure 1.4 Representations of the domain architectures of human p60 tim (transforming immortalized mammary oncogene), mouse βPix-c, human Vav oncogene, and human RGS1 (regulator of G-protein signaling 1). Domain abbreviations: C1, protein kinase C conserved region 1; C2, protein kinase C conserved region 2; CH, calponin homology; PH, pleckstrin homology; RasGAP, GTPase activator protein specific for Ras-like small GTPases; RhoGEF, guanine nucleotide exchange factor specific for Rho-type small GTPases; SH2, src homology 2; SH3, src homology 3. Species abbreviations: Hs, Homo sapiens; Mm, Mus musculus. [19]

1.2.2 The Properties of Protein Domains

Protein domains can fold independently from other parts of the protein, which makes them ideal building blocks in forming complicated proteins with novel functions. Members of homologous domain families can exist in different proteins or in different positions of the same protein. As we will discuss later, during evolution, domain swapping and domain combination are crucial ways of generating new proteins and obtaining new functions. Proteins sharing the same domain are evolutionarily related, and the relationship can be illustrated by tracing back through the evolutionary history of the common domains.

As spatially distinct parts of protein structures, protein domains possess all the characteristics of protein structures, and hence the structure of protein domains are more conserved than protein sequences. As a consequence, the number of possible domains and hence folds is much smaller than the number of protein sequences. The relatively vast number of possible sequences in the protein sequence universe corresponds to a small number of folds in the protein structure universe [4].

It's been widely accepted that the number of protein folds is limited. Initially the rate of discovery of novel folds deposited in the PDB each year was used to extrapolate and estimate the total protein structure universe [24, 25]. Recently, new methods that use the distribution of protein families within folds have been developed. Depending on the assumptions and methods used, the estimation of the total number of protein folds varies significantly from 1000 to 10000 [26-32].

Given that the total number of protein folds is limited, the structural genomics initiative [33, 34], like the human genome project that sequenced the entire human genome, aims to solve the 3-D structures covering all unique folds. Structural genomics groups select targets with no apparent sequence similarity to existing structures and hence have the highest likelihood to possess new folds. The ongoing structural genomics project intends to provide a complete protein structure space, with representative structures for all existing protein families, thereby enhancing our understanding of the structure and evolution of the protein universe.

Experimental results suggest that domains are fundamental functional units. Since protein structure determines protein function, a protein domain usually exhibits the same

or similar function in different structure contexts. For example, the protein kinase domain always has phosphorylation activity despite its adjacent substrate binding domains. Although some domains are not limited to one function, each protein domain is able to develop only a finite number of functions during evolution. Therefore, the protein structure universe as represented by all possible domains and their combinations corresponds to the protein functional space.

Different arrangements of protein domains not only generate new proteins, but also introduce new functions. The assignment of biochemical activity to a protein of unknown function is most straightforward when the new structure resembles that of the proteins whose functions are known. Although it is an over-simplification to consider the function of a protein as the sum of the function of its component domains, recognizing the domain architecture of a protein is usually the first step in investigating its function. For example, if an unknown protein contains a helix-turn-helix domain that typically binds to DNA, this protein is probably involved in transcription regulation or other related functions.

1.2.3 Protein Domain Classification

As the number of novel folds accumulates, a classification framework that depicts the internal relationship of protein domains is necessary. In the past ten years several schemes for protein structure classification have been developed and are widely used for studies of protein structure, function and evolution. All classifications attempt to build a

hierarchy of protein structures where the top levels of the hierarchy are defined by the three-dimensional structure and the lower levels are identified by sequence similarities. Here I will introduce and compare three of the major classifications, each differs in the method of curation, either manual or automated.

SCOP (Structural Classification Of Proteins) [http://scop.mrc-lmb.cam.ac.uk/scop/], developed and maintained by Murzin et al, is manually curated by experts in structural biology [35-37]. SCOP differs from other schemes by utilizing functional information to determine distant relationships when there is no obvious sequence similarity; therefore, SCOP is also considered an evolutionary classification. Although manual curation of SCOP means it is not up-to-date with the PDB, SCOP is considered the gold standard for protein structure classification. In my work, I chose to use SCOP among several potential classification schemes. Detailed information on SCOP will be given in Chapter 2.

CATH (Class, Architecture, Topology and Homologous superfamily) [http://www.cathdb.info/] was introduced by Thorton, Orengo and their colleagues [38-40]. CATH is constructed using a pipeline that combines manual and automated methods. Generally, computational methods are applied to most cases and human curation is used when discrepancies arise. The mixture of manual and automated methods ensures a balance between the speed and quality of classification.

FSSP database [http://www.ebi.ac.uk/dali/fssp], by Holm & Sander, provides fully automated fold classification based on the structure alignment of proteins [41]. Using the DALI structure alignment program, the database contains an all-against-all pair

wise structure alignment of available 3D structures, and groups proteins into clusters sharing similar folds. Since it is purely based on the 3D coordinates of proteins, FSSP is up-to-date with the PDB. The authors also provide a means for assigning homology despite low sequence similarity.

Although different in design and logic, the three classification schemes often agree. However, because there is no well established mathematical description of a protein's fold, i.e. the measurement of how much one fold is distinguished from others, sometimes the classification is quite arbitrary, and discrepancies between different classifications cannot be avoided. There are cases where different methods disagree in both the number of domains in a multidomain protein and the boundary of each domain. A comparison of the various classification schemes demonstrates that no classification scheme is perfect, especially when the domain assignment and structural similarity is ambiguous [42]. In my research, to ensure consistency, the SCOP classification is used entirely.

1.3 Protein Domain as Evolutionary Unit

1.3.1 The Power Law Distribution of Protein Domains

The protein structure classification scheme provides a direct view of the distribution of protein domains. The numbers of protein families in different protein folds are not equally distributed, as shown in Figure 1.5. Mathematically the distribution of protein folds by the number of constituent family members has been approximated by a power law, $y(x) \sim x^{-a}$, where y(x) is the frequency of folds that include exactly x family members, and a is a parameter typically between 1 and 3 [29]. The power law nature of protein folds indicates that a very small number of folds contain a lot of homologous proteins, whereas the majority of folds contain only one or a small number of homologs.

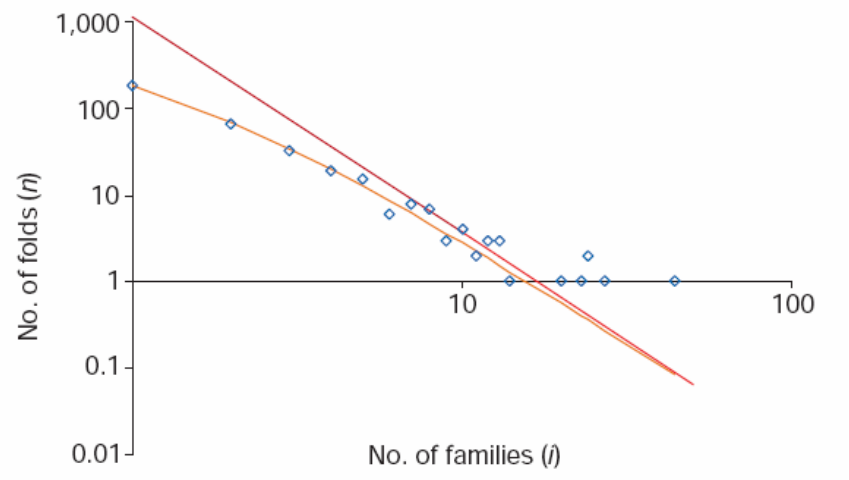


Figure 1.5 Double-logarithmic plot of the distribution of protein folds by the number of families [3].

A similar power law distribution has been observed in the distribution of the number of homologous domains in complete genomes [43, 44]. For example, the human genome contains 518 putative protein kinases [45], which regulate a variety of cellular functions, but has only one DNA gyrase, which regulates the number of topological links between two DNA strands.

The power law distribution, and the scale-free network derived from it, have also been found in other physical, biological and social contexts, such as protein – protein interaction networks, and the frequency of word usage in natural language. The common rationale behind the scale-free network is the preferential attachment during the evolution of the network, where the probability of a node acquiring new connections is proportional to the degree of the node [46-48].

The power law distribution of protein domains either in protein structure space or in individual genomes infers that protein domains are basic evolutionary units, and the distribution can only be understood by studying the evolution of each individual domain.

1.3.2 Protein Domain Analysis in the Era of Complete Genomes

The increasing numbers of complete genomes of various species across the whole tree of life provide biologists an immense amount of data never reached before. Comparative genomics aims at understanding the evolution and function of genomes and its components by the comparison of different genomes at various levels, such as DNA, protein, regulatory regions, etc. The genomic data, combined with powerful sequence-

based domain recognition algorithms, provide us the types and amount of protein domains in complete genomes [49, 50].

In 1997, Gerstein started to investigate the domain composition in the three superkingdoms of life using the standard sequence comparison method (FASTA) [51]. At that time, only one species from each superkingdom had been sequenced. Although FASTA has low performance in fold recognition (10~20%), the author found that "an appreciable fraction of the known folds are present in each of the major groups of organisms" and "most of the common folds are associated with many families." One year later, Gerstein compared the fold distribution from eight species, which were then clustered by their shared folds [52]. Zhang and Delisi, when estimating the total number of folds, illustrated the power law distribution of the family numbers in different folds [29]. In 1999, Wolf etc. improved the fold recognition method by using the profiled sequence comparison algorithm (Psi-Blast) and increased the genomic fold coverage to 20-30% [53]. Lin and Gerstein attempted to build a genomic tree based on the occurrence of protein folds and orthologs although only eight genomes were available in 2000 [54]. In 2001, Apic et al. not only counted the numbers of each individual fold in each genome, but also the domain combinations [55]. Using one of the most sensitive sequence comparison methods, the Hidden Markov Model, to study 40 complete genomes across archaea, bacteria and eukaryote, the authors demonstrated that the number of domain partners also follows a power law, and the combinations of existing domains are also important in inventing new proteins and new functions. Since then, many papers have been focused on the distribution [56, 57], combination [58], duplication [59-61],

divergence and convergence [62], fusion and fission [63, 64], and recombination [65, 66] of protein domains during evolution.

1.3.3 The History of Protein Domains

This previous work has convinced us that the protein domain is not only a structural concept, but also an evolutionary concept. In fact, protein domains are the fundamental evolutionary unit [19].

Protein domains are building blocks of proteins, but the same domain can exist in different proteins while still keeping its structure and function. Combinations of different domains and duplications of domains are the major evolutionary process in the acquisition of novel function. During evolution, protein domains can evolve by means of random mutation, and can be lost or horizontally transferred. Protein domains are more fundamental than entire proteins, and there is no smaller unit with which to study evolution.

All the evolutionary processes associated with protein domains that characterize them as evolutionary units are embedded in one basic physical property – the ability for independent folding. As evolutionary units, protein domains raise a series of interesting questions: How does the evolution of domains correlate with the evolution of genes, protein sequences, structures and functions? How many folds are there in the protein structure space? How many structure similarities arise through convergent evolution?

How did domains lose, transfer, and form new combinations over their entire evolutionary history? Finally, where did each domain come from?

As evolutionary units, how does the understanding of the evolution of protein domains contribute to the study of evolution of all forms of life? Can we use domain composition and distribution in genomes to characterize and classify organisms?

Addressing some of these questions is the subject of my dissertation. Chapter two describes the data source and algorithms used in this study. In chapter three, protein domain content is used to determine the relatedness of organisms and build the phylogenetic tree of life. In chapter four, whole genome alignments of closely-related species are performed using the positions of protein domains on each genome. In the last chapter, the evolutionary history of individual domains and domain combinations are discussed.

1.4 References

1. Holm, L. and C. Sander, *Mapping the protein universe.* Science, 1996. **273**(5275): p. 595-603.

2. Rost, B., *Did evolution leap to create the protein universe?* Curr Opin Struct Biol, 2002. **12**(3): p. 409-16.

3. Koonin, E.V., Y.I. Wolf, and G.P. Karev, *The structure of the protein universe and genome evolution.* Nature, 2002. **420**(6912): p. 218-23.

4. Chothia, C., et al., *Evolution of the protein repertoire.* Science, 2003. **300**(5626): p. 1701-3.

5. Dayhoff, M.O., *The origin and evolution of protein superfamilies.* Fed Proc, 1976. **35**(10): p. 2132-8.

6. Dayhoff, M.O., W.C. Barker, and L.T. Hunt, *Establishing homologies in protein sequences.* Methods Enzymol, 1983. **91**: p. 524-45.

7. Branden, C.-I. and J. Tooze, *Introduction to Protein Structure*. 1999, New York: Garland Publishing.

8. Ban, N., et al., *The complete atomic structure of the large ribosomal subunit at 2.4 A resolution.* Science, 2000. **289**(5481): p. 905-20.

9. Loll, B., et al., *Towards complete cofactor arrangement in the 3.0 A resolution structure of photosystem II.* Nature, 2005. **438**(7070): p. 1040-4.

10. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

11. Flaherty, K.M., et al., *Similarity of the three-dimensional structures of actin and the ATPase fragment of a 70-kDa heat shock cognate protein.* Proc Natl Acad Sci U S A, 1991. **88**(11): p. 5041-5.

12. Pastore, A. and A.M. Lesk, *Comparison of the structures of globins and phycocyanins: evidence for evolutionary relationship.* Proteins, 1990. **8**(2): p. 133-55.

13. Schnuchel, A., et al., *Structure in solution of the major cold-shock protein from Bacillus subtilis.* Nature, 1993. **364**(6433): p. 169-71.

14. Doolittle, R.F., *Of Urfs and Orfs: Primer on how to analyze derived amino acid sequences*. 1986, Mill Valley, CA, USA: University Science Books.

15. Doolittle, R.F., *The multiplicity of domains in proteins.* Annu Rev Biochem, 1995. **64**: p. 287-314.

16. Blake, C.C., et al., *Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Angstrom resolution.* Nature, 1965. **206**(986): p. 757-61.

17. Kartha, G., J. Bello, and D. Harker, *Tertiary structure of ribonuclease.* Nature, 1967. **213**(5079): p. 862-5.

18. Rossmann, M.G., D. Moras, and K.W. Olsen, *Chemical and biological evolution of nucleotide-binding protein.* Nature, 1974. **250**(463): p. 194-9.

19. Ponting, C.P. and R.R. Russell, *The natural history of protein domains.* Annu Rev Biophys Biomol Struct, 2002. **31**: p. 45-71.

20. Islam, S.A., J. Luo, and M.J. Sternberg, *Identification and analysis of domains in proteins.* Protein Eng, 1995. **8**(6): p. 513-25.

21. Siddiqui, A.S. and G.J. Barton, *Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions.* Protein Sci, 1995. **4**(5): p. 872-84.

22. Sowdhamini, R. and T.L. Blundell, *An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins.* Protein Sci, 1995. **4**(3): p. 506-20.

23. Swindells, M.B., *A procedure for detecting structural domains in proteins.* Protein Sci, 1995. **4**(1): p. 103-12.

24. Zuckerkandl, E., *The appearance of new structures and functions in proteins during evolution.* J Mol Evol, 1975. **7**(1): p. 1-57.

25. Chothia, C., *Proteins. One thousand families for the molecular biologist.* Nature, 1992. **357**(6379): p. 543-4.

26. Orengo, C.A., D.T. Jones, and J.M. Thornton, *Protein superfamilies and domain superfolds.* Nature, 1994. **372**(6507): p. 631-4.

27. Zhang, C.T., *Relations of the numbers of protein sequences, families and folds.* Protein Eng, 1997. **10**(7): p. 757-61.

28. Wang, Z.X., *A re-estimation for the total numbers of protein folds and superfamilies.* Protein Eng, 1998. **11**(8): p. 621-6.

29. Zhang, C. and C. DeLisi, *Estimating the number of protein folds.* J Mol Biol, 1998. **284**(5): p. 1301-5.

30. Govindarajan, S., R. Recabarren, and R.A. Goldstein, *Estimating the total number of protein folds.* Proteins, 1999. **35**(4): p. 408-14.

31. Wolf, Y.I., N.V. Grishin, and E.V. Koonin, *Estimating the number of protein folds and families from complete genome data.* J Mol Biol, 2000. **299**(4): p. 897-905.

32. Coulson, A.F. and J. Moult, *A unifold, mesofold, and superfold model of protein fold use.* Proteins, 2002. **46**(1): p. 61-71.

33. Burley, S.K., et al., *Structural genomics: beyond the human genome project.* Nat Genet, 1999. **23**(2): p. 151-7.

34. Chandonia, J.M. and S.E. Brenner, *The impact of structural genomics: expectations and outcomes.* Science, 2006. **311**(5759): p. 347-51.

35. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J Mol Biol, 1995. **247**(4): p. 536-40.

36. Murzin, A.G., *Structural classification of proteins: new superfamilies.* Curr Opin Struct Biol, 1996. **6**(3): p. 386-94.

37. Lo Conte, L., et al., *SCOP database in 2002: refinements accommodate structural genomics.* Nucleic Acids Res, 2002. **30**(1): p. 264-7.

38. Orengo, C.A., et al., *CATH--a hierarchic classification of protein domain structures.* Structure, 1997. **5**(8): p. 1093-108.

39. Todd, A.E., C.A. Orengo, and J.M. Thornton, *Evolution of function in protein superfamilies, from a structural perspective.* J Mol Biol, 2001. **307**(4): p. 1113-43.

40. Orengo, C.A., et al., *The CATH protein family database: a resource for structural and functional annotation of genomes.* Proteomics, 2002. **2**(1): p. 11-21.

41. Holm, L., et al., *A database of protein structure families with common folding motifs.* Protein Sci, 1992. **1**(12): p. 1691-8.

42. Veretnik, S., et al., *Toward consistent assignment of structural domains in proteins.* J Mol Biol, 2004. **339**(3): p. 647-78.

43. Huynen, M.A. and E. van Nimwegen, *The frequency distribution of gene family sizes in complete genomes.* Mol Biol Evol, 1998. **15**(5): p. 583-9.

44. Qian, J., N.M. Luscombe, and M. Gerstein, *Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model.* J Mol Biol, 2001. **313**(4): p. 673-81.

45. Manning, G., et al., *The protein kinase complement of the human genome.* Science, 2002. **298**(5600): p. 1912-34.

46. Barabasi, A.L., *Linked: The New Science of Networks*. 2002, New York: Perseus.

47. Barabasi, A.L. and R. Albert, *Emergence of scaling in random networks.* Science, 1999. **286**(5439): p. 509-12.

48. Bilke, S. and C. Peterson, *Topological properties of citation and metabolic networks.* Phys Rev E Stat Nonlin Soft Matter Phys, 2001. **64**(3 Pt 2): p. 036106.

49. Copley, R.R., et al., *Protein domain analysis in the era of complete genomes.* FEBS Lett, 2002. **513**(1): p. 129-34.

50. Lee, D., et al., *A structural perspective on genome evolution.* Curr Opin Struct Biol, 2003. **13**(3): p. 359-69.

51. Gerstein, M., *A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure.* J Mol Biol, 1997. **274**(4): p. 562-76.

52. Gerstein, M., *Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census.* Proteins, 1998. **33**(4): p. 518-34.

53. Wolf, Y.I., et al., *Distribution of protein folds in the three superkingdoms of life.* Genome Res, 1999. **9**(1): p. 17-26.

54. Lin, J. and M. Gerstein, *Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels.* Genome Res, 2000. **10**(6): p. 808-18.

55. Apic, G., J. Gough, and S.A. Teichmann, *Domain combinations in archaeal, eubacterial and eukaryotic proteomes.* J Mol Biol, 2001. **310**(2): p. 311-25.

56. Ranea, J.A., et al., *Evolution of protein superfamilies and bacterial genome size.* J Mol Biol, 2004. **336**(4): p. 871-87.

57. Vogel, C., et al., *Supra-domains: evolutionary units larger than single protein domains.* J Mol Biol, 2004. **336**(3): p. 809-23.

58. Ekman, D., et al., *Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions.* J Mol Biol, 2005. **348**(1): p. 231-43.

59. Vogel, C. and V. Morea, *Duplication, divergence and formation of novel protein topologies.* Bioessays, 2006. **28**(10): p. 973-8.

60. Vogel, C. and C. Chothia, *Protein family expansions and biological complexity.* PLoS Comput Biol, 2006. **2**(5): p. e48.

61. Bjorklund, A.K., D. Ekman, and A. Elofsson, *Expansion of protein domain repeats.* PLoS Comput Biol, 2006. **2**(8): p. e114.

62. Gough, J., *Convergent evolution of domain architectures (is rare).* Bioinformatics, 2005. **21**(8): p. 1464-71.

63. Kummerfeld, S.K. and S.A. Teichmann, *Relative rates of gene fusion and fission in multi-domain proteins.* Trends Genet, 2005. **21**(1): p. 25-30.

64. Pasek, S., J.L. Risler, and P. Brezellec, *Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins.* Bioinformatics, 2006. **22**(12): p. 1418-23.

65. Vogel, C., S.A. Teichmann, and J. Pereira-Leal, *The relationship between domain duplication and recombination.* J Mol Biol, 2005. **346**(1): p. 355-65.

66. Bjorklund, A.K., et al., *Domain rearrangements in protein evolution.* J Mol Biol, 2005. **353**(4): p. 911-23.

Chapter 2 Data and Algorithms

2.1 Databases and Algorithms

2.1.1 PDB Data

The Protein Data Bank (PDB) [1] is a public repository of the 3D structures of macromolecules solved mainly by X-ray crystallography and Nuclear Magnetic Resonance (NMR). As of Dec 19 2006, there were 40479 structures in the PDB and the number increases by about 150 per week. The PDB file for each structure not only stores the primary atomic Cartesian coordinates of the protein and other molecules in the structure, but also contains other information (metadata) related to that protein structure.

For example, disulfide bond information can be retrieved from PDB files when the author labeled them in the file when creating the PDB entry. The SOURCE records in PDB files usually describe the organism from which the protein structures come from. Because the formats of PDB files are not uniform, and sometimes proteins are synthetic proteins that come from multiple sources, a data unification step is needed to obtain the correct information.

2.1.2 SCOP Protein Structure Classification

In recent decades, the numbers of three-dimensional structures of proteins have been accumulating rapidly. Based on the experimental data, the Structure Classification

Of Proteins (SCOP) [2, 3] aims to build a hierarchical classification that describes the structural and evolutionary relationship of proteins. The SCOP is manually defined by experts in structural biology, and therefore is considered the gold standard of protein structure classification. The four levels in SCOP, from bottom to top, are:

FAMILY: Protein domains are clustered together into families if all domains have sequence identities of 30% and greater. SCOP families contain homologous domains that have a common evolutionary origin.

SUPERFAMILY: Families, whose proteins have low sequence identities but their structures and, in many cases, functional features suggest that a common evolutionary origin is probable, are placed together in superfamilies; for example, actin, the ATPase domain of the heat-shock protein and hexokinase are three families that form a superfamily [4].

FOLD: Superfamilies are defined as having a common fold if their proteins have same major secondary structures in the same arrangement with the same topological connections.

CLASS: Folds are grouped into classes on the basis of the secondary structures of which they composed: (1) all alpha (for proteins whose structure is essentially formed by α-helices), (2) all beta (for those whose structure is essentially formed by β-sheets), (3) alpha / beta (for proteins with α-helices and β-strands that are largely interspersed), (4) alpha + beta (for those in which α-helices and β-strands are largely segregated), (5) multi-domain (for those with domains of different fold and for which no homologues

are known at present), (6) membrane and cell surface proteins and (7) small proteins whose stability is enhanced by disulphide bond or other means.

The Class level is defined for user convenience. The Fold level represents the structural similarity, which may arise from the physical and chemical properties of a protein that prefers a certain packing arrangement and chain topology. The structural similarity may result either from divergence of distantly related proteins or convergence of evolutionarily unrelated proteins. Only if additional functional relatedness that suggests common ancestry can families with common fold be clustered into superfamilies. In other words, families with similar folds may evolve from a common ancestor, but due to the lack of evolutionary evidence, their relationship has not yet been recovered.

Many protein structures contain multiple domains. SCOP first cuts these structures into individual domains, according to either the principle of structural compactness or recurrence in other proteins (class e). Although some proteins apparently exhibit two spatially distinct parts, if their components are not found in any other proteins, the whole protein is considered as one domain. For instance, the protein kinase contains an N-terminal domain and a C-terminal domain, but they are always together and the phosphorylation function of the protein kinase requires the cooperation of both domains. Therefore, in SCOP, the whole protein is classified as one the protein kinase domain.

As the PDB grows, the content of SCOP also changes, as shown in Table 2.1.

Table 2.1 SCOP Classification Statistics

	1995	1.55 (2002)	1.65 (Dec 2003)	1.67 (Feb 2005)	1.69 (July 2005)
Class	5	7	7	7	7
Fold	279	605	800	887	945
Superfamily	366	947	1294	1447	1539
Family	498	1557	2327	2630	2845
Domain	3179	30403	54745	65122	70859

2.1.3 Complete Genome Sequences

By 2004, over 200 species have been completely sequenced, including over 100 bacteria, more than ten Archaea and thirty Eukaryotes from most major lineages. The increasing genomic data from all major branches in the phylogenetic tree provide a plentiful resource for comparative genomics and the study of evolution [5].

In this study, protein domain assignment at the genomic level requires complete proteome data, which are predicted from the genome data through the gene finding step [6]. Along with genetic mapping, gene finding is the first step in genome annotation at the nucleic acid level. In small prokaryotic genomes, gene finding is mainly identifying long open reading frames (ORFs). As genomes get larger, gene finding become increasingly tricky, because of the lower signal-to-noise ratio (ratio of coding region vs. non-coding region), the complexity in gene boundary and exon boundary and the variety of alterative splicing forms. Lots of computational gene finding program have been

developed based on three major approaches: *ab initio* approach based on genomic DNA sequence alone, extrinsic approach that finds the sequence similarity with the known genes, cDNAs, and proteins etc., and the comparative genomics approach that is based on comparison of the complete genome of different species. Gene finding that combines several approaches gets better accuracy and coverage than any single algorithm.

Genome sequencing is becoming more convenient and reliable, but sequencing errors still exist. For example, based on our analysis, the genome of the frog *Xenopus tropicalis* is contaminated by substantial bacteria genetic materials [7], such data are misleading when studying the evolution of protein domains and organisms. Moreover, since new experimental data keep coming out that correct and improve gene prediction, complete whole genome gene prediction is an endeavor that is constantly improving. Notwithstanding, missing predictions and errors in gene prediction are inevitable, and will be part of the errors in our domain assignment data.

2.1.4 Hidden Markov Model

Profile Hidden Markov Models (HMM) have been demonstrated to be very effective in detecting conserved patterns in multiple sequences [8, 9]. The typical HMM is a chain of match (square), insert (diamond) and delete (circle) nodes, with all transitions between nodes and all character costs in the insert and match nodes trained to specific probabilities (Figure 2.1). The single best path through an HMM corresponds to a path from the Start state to the End state in which each character of the sequence is related to a

successive match or insertion state along that path (delete states indicate that the sequence has no character corresponding to that position in the HMM). A study that compared the performance of prediction of protein similarity of HMM (SAM-T98) [10] with the standard methods BLAST, FASTA and another profile method PSI-BLAST and ISS showed that HMM is superior to PSI-BLAST, which is superior to ISS, which is superior to BLAST and FASTA [11].

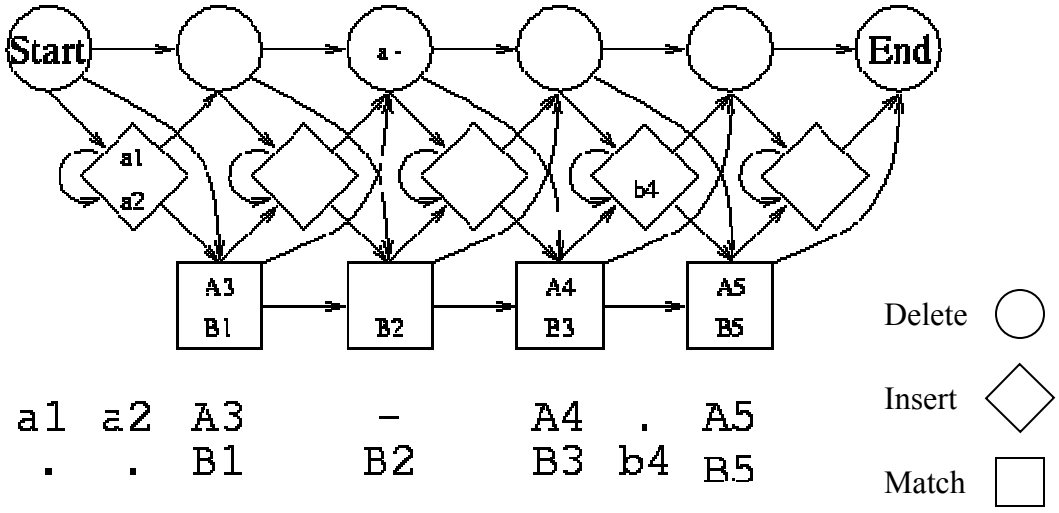


Figure 2.1 A linear hidden Markov model and example alignment

HMM is by far the most sensitive homology detection algorithm using multiple sequence alignment, so it is the best choice for the high-throughput genomic protein domain assignment.

2.1.5 SUPERFAMILY Database

The SUPERFAMILY database [12] contains the genomic protein domain assignment based on the SCOP structural classification of protein domains using the profile HMM multiple sequence alignment algorithm. Initially, SUPERFAMILY was built at the SCOP superfamily level, within which structural, functional and sequence evidences suggest protein domains have common evolutionary ancestor. In the latest release, it also extends its coverage to the family level that represents closely related homologs. The database uses hidden Markov models (HMMs), which are profiles based on multiple sequence alignments designed to represent a protein family (or superfamily) and can be used to search genomic sequence databases for homologues.

Domain assignment in the SUPERFAMILY database intrinsically contains evolutionary relationships, because the SCOP superfamily level denotes the remote homologous relationship that cannot be derived by sequence information alone.

SUPERFAMILY database is updated according to the new release of SCOP and the availability of newly sequenced genomes. As of July 2004, there are 17 Archaea, 154 Bacteria and 38 Eukaryote species with complete genomes analyzed and deposited in the database, and the number has kept growing. Table 2.2 lists the number of complete genomes and SCOP version in each release of SUPERFAMILY database. During my Ph.D. study, two releases of the SUPERFAMILY data were applied at different stages in my research. The lists of all species in the two releases of SUPERFAMILY database are shown in Tables 2.3 and 2.4.

Table 2.2 The SCOP version and number of annotated complete genomes in each release of the SUPERFAMILY database. Some complete genomes belong to different strains of the same species. The number of unique genomes in each category only considers a single strain of any given species.

SUPERFAMILY Release		2004, 07, 25	2005, 10, 09
SCOP Version		1.65	1.69
Archaea	Total	19	23
	Unique	19	23
Bacteria	Total	154	233
	Unique	119	172
Eukaryote	Total	38	59
	Unique	36	59
Life	Total	212	315
	Unique	174	254

On average, the domain assignment in the database covers about 40% ~ 60% of the proteins in each genome, which is significantly higher than in works done by other methods. This coverage is expected to increase as structural genomics projects solve more novel structures, giving a more complete structural picture of the genomes.

Other databases aiming at the same purpose includes PEDANT [13] and Gene3D [14]. PEDANT is a comprehensive genome database that provides automatic analysis of genomic sequences by a large variety of established bioinformatics tools, but its domain assignment is not as consistent and up-to-date as SUPERFAMILY. Gene3D closely resembles SUPERFAMILY; the only difference is that Gene3D is based on the CATH domain classification instead of SCOP. Since the classification and coverage of SCOP and CATH is similar, the results of Gene3D and SUPERFAMILY are often comparable.

In my study, I retrieve the data from SUPERFAMILY exclusively. However, because the data are from the public database and are handled using standard bioinformatics methods, such as the domain assignment, my results and conclusions are not limited to the SUPERFAMILY database and should be universal.

2.2 Lists of Organisms in This Study

2.2.1 SUPERFAMILY Release 2004-07-25

Table 2.3 The complete genomes in SUPERFAMILY database release 2004-07-25.

A) Eukaryote species list

	Genome Name	Fold*	Superfamily*
Agam	Anopheles gambiae 22.2b	544	807
Agos	Ashbya gossypii 1.0	413	593
Anid	Aspergillus nidulans 1 r3.1	460	678
Atha	Arabidopsis thaliana 5	519	773
Calb	Candida albicans	422	602
Cbri	Caenorhabditis briggsae Aug03	508	745
Cele	Caenorhabditis elegans WS123	523	769
Cgla	Candida glabrata	411	580
Cint	Ciona intestinalis 1.0	518	762
Ddis	Dictyostelium discoideum 2	451	654
Dhan	Debaromyces hansenii	429	607
Dmel	Drosophila melanogaster 3.2	505	750
Drer	Danio rerio 22.3b	525	771
Ecun	Encephalitozoon cuniculi	256	326
Fgra	Fusarium graminearum 1	475	706
Frub	Fugu rubripes 22.2c	541	804
Ggal	Gallus gallus 22.1	534	792
Hsap	Homo sapiens 22.34d	567	862
Klac	Kluyveromyces lactis	414	589
Kwal	Kluyveromyces waltii	412	586
Mgri	Magnaporthe grisea 7 r2.3	457	678
Mmus	Mus musculus 22.32b	572	870
Ncra	Neurospora crassa 3	472	697
Osat	Oryza sativa ssp. japonica 2.0	515	777
Pfal	Plasmodium falciparum 1	362	501

Table 2.3 (A) continued

	Genome Name	Fold*	Superfamily*
Ptro	Pan troglodytes 22.1	558	837
Pyoe	Plasmodium yoelii ssp. yoelii 1	374	512
Rnor	Rattus norvegicus 22.3b	565	852
Sbay	Saccharomyces bayanus MIT	428	609
Scer	Saccharomyces cerevisiae	439	631
Smik	Saccharomyces mikatae MIT	426	605
Spar	Saccharomyces paradoxus MIT	428	614
Spom	Schizosaccharomyces pombe	428	618
Umay	Ustilago maydis 1 r2	418	610
Xtro	Xenopus tropicalis 2.0	589	889
Ylip	Yarrowia lipolytica	417	603
Strains*	Genome Name	Folds	Superfamilies
Osat1	Oryza sativa ssp. indica	530	840
Tbru	Trypanosoma brucei	69	78

*Fold/Superfamily, the different types of fold/FSF assigned to each complete genome

*Strains, the organisms belong to different strains of the same species

B) Archaea species list

	Genome Name	Mb*	Fold	Superfamily
Aful	Archaeoglobus fulgidus DSM 4304	2.18	307	434
Aper	Aeropyrum pernix	1.67	271	368
Halsp	Halobacterium sp. NRC-1	2.01	296	415
Mace	Methanosarcina acetivorans C2A	5.75	339	496
Mjan	Methanocaldococcus jannaschii	1.66	268	384
Mkan	Methanopyrus kandleri AV19	1.69	262	358

Table 2.3 (B) continued

	Genome Name	Mb*	Fold	Superfamily
Mmar	Methanococcus maripaludis	1.66	274	390
Mmaz	Methanosarcina mazei Goe1	4.10	330	479
Mthe	Methanothermobacter thermautotrophicus Delta H	1.75	292	417
Nequ	Nanoarchaeum equitans Kin4-M	0.49	144	180
Paby	Pyrococcus abyssi	1.77	281	403
Pfur	Pyrococcus furiosus DSM 3638	1.91	289	414
Phor	Pyrococcus horikoshii	1.74	275	391
Ptor	Picrophilus torridus DSM 9790	1.55	281	376
PYaer	Pyrobaculum aerophilum IM2	2.22	291	398
Ssol	Sulfolobus solfataricus	2.99	303	427
Stok	Sulfolobus tokodaii	2.69	299	416
Taci	Thermoplasma acidophilum	1.56	278	384
Tvol	Thermoplasma volcanium	1.58	278	382

*Mb, Genome length in Megabase

C) Bacteria species list

	Genome Name	Mb	Fold	Superfamily
Aaeo	Aquifex aeolicus VF5	1.55	327	459
Acisp	Acinetobacter sp. ADP1	3.60	372	538
Atum	Agrobacterium tumefaciens C58	2.84	413	592
Bant	Bacillus anthracis Ames	5.23	408	592
Baph	Buchnera aphidicola Bp	0.62	240	309
Bbac	Bdellovibrio bacteriovorus HD100	3.78	357	513
Bbro	Bordetella bronchiseptica RB50	5.34	408	590
Bbur	Borrelia burgdorferi B31	0.91	228	299
Bcer	Bacillus cereus ATCC 14579	5.41	412	601
Bhal	Bacillus halodurans	4.20	403	580

Table 2.3 (C) continued

	Genome Name	Mb	Fold	Superfamily
Bhen	Bartonella henselae Houston-1	1.93	310	419
Bjap	Bradyrhizobium japonicum USDA 110	9.11	429	623
Blon	Bifidobacterium longum NCC2705	2.26	307	430
Bmel	Brucella melitensis 16M	3.29	385	550
Bpar	Bordetella parapertussis 12822	4.77	404	583
Bper	Bordetella pertussis Tohama I	4.09	384	556
Bqui	Bartonella quintana	1.58	300	404
Bsub	Bacillus subtilis ssp. subtilis 168	4.21	402	585
Bsui	Brucella suis 1330	3.31	383	544
Bthe	Bacteroides thetaiotaomicron VPI-5482	6.26	355	510
Bthu	Bacillus thuringiensis ser. konkukian 97-27	5.24	405	592
Cace	Clostridium acetobutylicum	3.94	378	547
Cblo	Candidatus Blochmannia floridanus	0.71	253	334
Cbur	Coxiella burnetii RSA 493	2.00	333	465
Ccav	Chlamydophila caviae GPIC	1.17	257	337
Ccre	Caulobacter crescentus CB15	4.02	395	570
Cdip	Corynebacterium diphtheriae	2.49	348	493
Ceff	Corynebacterium efficiens YS-314	3.15	358	512
Cglu	Corynebacterium glutamicum ATCC 13032	3.31	362	509
Cjej	Campylobacter jejuni ssp. jejuni NCTC 11168	1.64	315	448
Cmur	Chlamydia muridarum	1.07	249	327
Cper	Clostridium perfringens 13	3.03	365	518
Cpne	Chlamydophila pneumoniae J138	1.23	250	328
Ctep	Chlorobium tepidum TLS	2.15	345	492
Ctet	Clostridium tetani E88	2.80	338	476
Ctre	Chlamydia trachomatis	1.04	248	322
Cvio	Chromobacterium violaceum ATCC 12472	4.75	420	598
Drad	Deinococcus radiodurans R1	3.06	371	527
Dvul	Desulfovibrio vulgaris ssp. vulgaris Hildenborough	3.57	351	496
Ecar	Erwinia carotovora ssp. atroseptica SCRI1043	5.06	417	616
Ecol	Escherichia coli K12	4.64	440	647
Efae	Enterococcus faecalis V583	3.22	363	504

Table 2.3 (C) continued

	Genome Name	Mb	Fold	Superfamily
Fnuc	Fusobacterium nucleatum ssp. nucleatum ATCC 25586	2.17	319	450
Gsul	Geobacter sulfurreducens PCA	3.81	383	554
Gvio	Gloeobacter violaceus	4.66	389	572
Hduc	Haemophilus ducreyi 35000HP	1.70	329	454
Hhep	Helicobacter hepaticus ATCC 51449	1.80	322	453
Hinf	Haemophilus influenzae Rd KW20	1.83	345	483
Hpyl	Helicobacter pylori 26695	1.67	303	418
Linn	Listeria innocua	3.01	361	516
Lint	Leptospira interrogans ser. lai 56601	4.69	356	509
Ljoh	Lactobacillus johnsonii NCC 533	1.99	298	394
Llac	Lactococcus lactis ssp. lactis	2.37	330	455
Lmon	Listeria monocytogenes EGD-e	2.94	364	519
Lpla	Lactobacillus plantarum WCFS1	3.31	340	468
Mavi	Mycobacterium avium ssp. paratuberculosis k10	4.83	370	528
Mbov	Mycobacterium bovis AF2122/97	4.35	375	540
Mgal	Mycoplasma gallisepticum R	1.00	199	249
Mgen	Mycoplasma genitalium	0.58	183	234
Mlep	Mycobacterium leprae	3.27	317	447
Mlot	Mesorhizobium loti	7.04	424	609
Mmob	Mycoplasma mobile 163K	0.78	189	237
Mmyc	Mycoplasma mycoides ssp. mycoides SC	1.21	207	263
Mpen	Mycoplasma penetrans	1.36	213	281
Mpne	Mycoplasma pneumoniae	0.82	192	243
Mpul	Mycoplasma pulmonis	0.96	197	250
Mtub	Mycobacterium tuberculosis H37Rv	4.41	377	541
Neur	Nitrosomonas europaea ATCC 19718	2.81	371	531
Nmen	Neisseria meningitidis MC58	2.27	345	488
Nossp	Nostoc sp. PCC 7120	6.41	403	585
Oihe	Oceanobacillus iheyensis HTE831	3.63	382	539
OYPH	Onion yellows phytoplasma	0.86	174	219
Parsp	Parachlamydia sp. UWE25	2.41	308	415

Table 2.3 (C) continued

	Genome Name	Mb	Fold	Superfamily
Pgin	Porphyromonas gingivalis W83	2.34	303	422
Pirsp	Pirellula sp. 1	7.14	384	551
Plum	Photorhabdus luminescens ssp. laumondii TTO1	5.69	422	613
Pmar	Prochlorococcus marinus ssp. marinus CCMP1375	1.75	315	444
Pmul	Pasteurella multocida	2.26	372	526
Pput	Pseudomonas putida KT2440	6.18	436	633
PSaer	Pseudomonas aeruginosa PAO1	6.26	442	649
Psyr	Pseudomonas syringae pv. tomato DC3000	6.40	426	625
Rcon	Rickettsia conorii	1.27	280	368
Rpal	Rhodopseudomonas palustris CGA009	5.46	412	606
Rpro	Rickettsia prowazekii	1.11	268	346
Rsol	Ralstonia solanacearum	3.71	403	584
Rsph	Rhodobacter sphaeroides		409	588
Saga	Streptococcus agalactiae 2603V/R	2.16	327	447
Saur	Staphylococcus aureus ssp. aureus Mu50	2.88	371	517
Save	Streptomyces avermitilis MA-4680	9.03	409	609
Scoe	Streptomyces coelicolor A3(2)	8.67	412	605
Sent	Salmonella enterica ssp. enterica ser. Typhi	4.81	433	640
Sepi	Staphylococcus epidermidis ATCC 12228	4.50	361	500
Sfle	Shigella flexneri 2a 301	4.61	430	629
Smel	Sinorhizobium meliloti	3.65	388	555
Smut	Streptococcus mutans UA159	2.03	325	451
Sone	Shewanella oneidensis MR-1	4.97	420	608
Spne	Streptococcus pneumoniae TIGR4	2.16	322	442
Spyo	Streptococcus pyogenes M1 GAS	1.85	321	442
Styp	Salmonella typhimurium LT2	4.86	440	648
Syn1	Synechococcus sp. WH 8102	2.43	352	504
Syn2	Synechocystis sp. PCC 6803	3.57	378	542
Tden	Treponema denticola ATCC 35405	2.84	307	435
Telo	Thermosynechococcus elongatus BP-1	2.59	360	518
Tmar	Thermotoga maritima	1.86	320	461
Tpal	Treponema pallidum ssp. pallidum Nichols	1.14	251	336

Table 2.3 (C) continued

	Genome Name	Mb	Fold	Superfamily
Tten	Thermoanaerobacter tengcongensis	2.69	347	497
Tthe	Thermus thermophilus HB27	1.89	335	478
Twhi	Tropheryma whipplei Twist	0.93	252	334
Uure	Ureaplasma urealyticum	0.75	182	233
Vcho	Vibrio cholerae O1 biovar eltor N16961	4.03	410	595
Vpar	Vibrio parahaemolyticus RIMD 2210633	5.15	424	617
Vvul	Vibrio vulnificus CMCP6	5.12	421	613
Wglo	Wigglesworthia glossinidia	0.70	248	341
Wolb	Wolbachia	1.27	279	369
Wsuc	Wolinella succinogenes	2.11	321	462
Xaxo	Xanthomonas axonopodis pv. citri 306	5.17	421	617
Xcam	Xanthomonas campestris pv. campestris ATCC 33913	5.08	415	611
Xfas	Xylella fastidiosa 9a5c	2.68	362	518
Ypes	Yersinia pestis CO92	4.65	425	616
Strains	Names		Folds	Superfamilies
Saur1	Staphylococcus aureus ssp. aureus MRSA252		372	519
Saur2	Staphylococcus aureus ssp. aureus MSSA476		373	519
Bant4	Bacillus anthracis Sterne		410	598
Bant1	Bacillus anthracis A2012		401	586
Bant3	Bacillus anthracis Ames 0581		405	587
Baph3	Buchnera aphidicola Sg		244	322
Baph1	Buchnera aphidicola APS		253	334
Bcer1	Bacillus cereus ATCC 10987		404	591
Cpne4	Chlamydophila pneumoniae TW-183		249	327
Cpne1	Chlamydophila pneumoniae AR39		249	325
Cpne2	Chlamydophila pneumoniae CWL029		249	326
Ecol1	Escherichia coli CFT073		439	640
Ecol4	Escherichia coli O157:H7 EDL933		449	656
Ecol3	Escherichia coli O157:H7		450	657
Sent2	Salmonella enterica ssp. enterica ser. Typhi Ty2		433	640

Table 2.3 (C) continued

	Genome Name	Mb	Fold	Superfamily
Lint1	Leptospira interrogans ser. Copenhageni Fiocruz L1-130		355	507
Sfle1	Shigella flexneri 2a 2457T		425	620
Hpyl2	Helicobacter pylori J99		296	408
Lmon1	Listeria monocytogenes 4b F2365		360	513
Mtub1	Mycobacterium tuberculosis CDC1551		376	541
Nmen2	Neisseria meningitidis Z2491		348	489
Pmar1	Prochlorococcus marinus MIT 9313		336	476
Pmar3	Prochlorococcus marinus ssp. pastoris CCMP1986		325	459
Saur4	Staphylococcus aureus ssp. aureus MW2		377	522
Saur5	Staphylococcus aureus ssp. aureus N315		371	517
Saga2	Streptococcus agalactiae NEM316		372	449
Spyo2	Streptococcus pyogenes MGAS315		320	440
Spyo4	Streptococcus pyogenes SSI-1		319	438
Spne1	Streptococcus pneumoniae R6		325	448
Spyo3	Streptococcus pyogenes MGAS8232		319	442
Twhi1	Tropheryma whipplei TW08/27		248	331
Xfas2	Xylella fastidiosa Temecula1		358	519
Ypes1	Yersinia pestis biovar Mediaevails 91001		425	616
Ypes3	Yersinia pestis KIM		424	614

2.2.2 SUPERFAMILY Release 2005-10-09

The additional organisms curated since the last release are listed in the following tables.

Table 2.4 The genomes added in the SUPERFAMILY since release 2004-07-25

A) Eukaryote species list

Eukaryote	Genome Name
Amel	Apis mellifera 32.2a
Afum	Aspergillus fumigatus
Btau	Bos taurus 32.1
Cneo	Cryptococcus neoformans 1
Crei	Chlamydomonas reinhardtii 2.0
Cfam	Canis familiaris 32.1d
Dpse	Drosophila pseudoobscura 1.04
Lmaj	Leishmania major 100505
Ehis	Entamoeba histolytica 1
Pchr	Phanerochaete chrysosporium 2.1
TEthe	Tetrahymena thermophila 1
Snod	Stagonospora nodorum 1
Tann	Theileria annulata
Tpar	Theileria parva 1
Pram	Phytophthora ramorum 1.0
Tree	Trichoderma reesei 1.2
Chom	Cryptosporidium hominis
Psoj	Phytophthora sojae 1.0
Tpse	Thalassiosira pseudonana
Tnig	Tetraodon nigroviridis 32.1d
Tcru	Trypanosoma cruzi
Cmer	Cyanidioschyzon merolae

B) Archaea species list

Archaea	Genome Name
Mbar	Methanosarcina barkeri fusaro
Hmar	Haloarcula marismortui ATCC 43049
Mmar	Methanococcus maripaludis S2
Tkod	Thermococcus kodakaraensis KOD1
Saci	Sulfolobus acidocaldarius DSM 639

C) Bacteria species list

Bacteria	Genome Name
Pacn	Propionibacterium acnes KPA171202
Amar	Anaplasma marginale St. Maries
Azosp	Azoarcus sp. EbN1
Babo	Brucella abortus biov. 1 9-941
Bmal	Burkholderia mallei ATCC 23344
Cglu1	Corynebacterium glutamicum ATCC 13032
Cabo	Chlamydophila abortus S26/3
Cglu2	Corynebacterium glutamicum ATCC 13032
Daro	Dechloromonas aromatica RCB
Dehsp	Dehalococcoides sp. CBDB1
Dpsy	Desulfotalea psychrophila LSv54
Erum3	Ehrlichia ruminantium Gardel
Erum2	Ehrlichia ruminantium Welgevonden
Erum1	Ehrlichia ruminantium Welgevonden
Ecan	Ehrlichia canis Jake
Sepi2	Staphylococcus epidermidis RP62A
Deth	Dehalococcoides ethenogenes 195
Bfra2	Bacteroides fragilis YCH46
Bfra1	Bacteroides fragilis NCTC 9343
Ftul	Francisella tularensis ssp. tularensis SCHU S4
Bgar	Borrelia garinii PBi

Table 2.4 (C) continued

Bacteria	Genome Name
Gkau	Geobacillus kaustophilus HTA426
Ppro	Photobacterium profundum SS9
Hinf2	Haemophilus influenzae 86-028NP
Mhyo2	Mycoplasma hyopneumoniae 7448
Iloi	Idiomarina loihiensis L2TR
Cjej2	Campylobacter jejuni RM1221
Cjei	Corynebacterium jeikeium K411
Blic1	Bacillus licheniformis ATCC 14580
Lpne1	Legionella pneumophila Paris
Blic2	Bacillus licheniformis ATCC 14580
Lpne2	Legionella pneumophila Lens
Lint2	Leptospira interrogans ser. Lai 56601
Laci	Lactobacillus acidophilus NCFM
Cpsy	Colwellia psychrerythraea 34H
Lxsp	Leifsonia xyli ssp. xyli CTCB07
Msuc	Mannheimia succiniciproducens MBEL55E
Mflo	Mesoplasma florum L1
Mcap	Methylococcus capsulatus Bath
Mhyo3	Mycoplasma hyopneumoniae J
Nfar	Nocardia farcinica IFM 10152
Ngon	Neisseria gonorrhoeae FA 1090
THden	Thiobacillus denitrificans ATCC 25259
Cblo2	Candidatus Blochmannia pennsylvanicus BPEN
Pmar4	Prochlorococcus marinus NATL2A
Sson	Shigella sonnei Ss046
Goxy	Gluconobacter oxydans 621H
Bpse	Burkholderia pseudomallei K96243
Pflu	Pseudomonas fluorescens Pf-5
Psyr2	Pseudomonas syringae pv. phaseolicola 1448A
Rfel	Rickettsia felis URRWXCal2
Parc	Psychrobacter arcticum 273-4
Reut	Ralstonia eutropha JMP134
Rtyp	Rickettsia typhi Wilmington

Table 2.4 (C) continued

Bacteria	Genome Name
SYthe	Symbiobacterium thermophilum IAM 14863
STthe1	Streptococcus thermophilus LMG 18311
SIpom	Silicibacter pomeroyi DSS-3
Spyo5	Streptococcus pyogenes MGAS10394
Tfus	Thermobifida fusca YX
THthe2	Thermus thermophilus HB8
Cpel	Candidatus Pelagibacter ubique HTCC1062
Psyr3	Pseudomonas syringae pv. syringae B728a
Vfis	Vibrio fischeri ES114
Nwin	Nitrobacter winogradskyi Nb-255
Wend	Wolbachia endosymbiont TRS of Brugia malayi
Sent4	Salmonella enterica ssp. enterica ser. Paratyphi A ATCC
STthe2	Streptococcus thermophilus CNRZ1066
Spyo1	Streptococcus pyogenes M1 GAS
Selo	Synechococcus elongatus PCC 6301
Bcla	Bacillus clausii KSM-K16
Bcer3	Bacillus cereus E33L
Sent3	Salmonella enterica ssp. enterica ser. Choleraesuis
Lpne3	Legionella pneumophila ssp. pneumophila Philadelphia 1
Saur6	Staphylococcus aureus ssp. aureus COL
Mhyo1	Mycoplasma hyopneumoniae 232
Xcam2	Xanthomonas campestris pv. campestris 8004
Xory	Xanthomonas oryzae pv. oryzae KACC10331
Msyn	Mycoplasma synoviae 53

2.3 References

1. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

2. Murzin, A.G., *Structural classification of proteins: new superfamilies.* Curr Opin Struct Biol, 1996. **6**(3): p. 386-94.

3. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J Mol Biol, 1995. **247**(4): p. 536-40.

4. Flaherty, K.M., et al., *Similarity of the three-dimensional structures of actin and the ATPase fragment of a 70-kDa heat shock cognate protein.* Proc Natl Acad Sci U S A, 1991. **88**(11): p. 5041-5.

5. Miller, W., et al., *Comparative genomics.* Annu Rev Genomics Hum Genet, 2004. **5**: p. 15-56.

6. Stein, L., *Genome annotation: from sequence to biology.* Nat Rev Genet, 2001. **2**(7): p. 493-503.

7. Yang, S., R.F. Doolittle, and P.E. Bourne, *Phylogeny determined by protein domain content.* Proc Natl Acad Sci U S A, 2005. **102**(2): p. 373-8.

8. Krogh, A., et al., *Hidden Markov models in computational biology. Applications to protein modeling.* J Mol Biol, 1994. **235**(5): p. 1501-31.

9. Eddy, S.R., *Multiple alignment using hidden Markov models.* Proc Int Conf Intell Syst Mol Biol, 1995. **3**: p. 114-20.

10. Karplus, K., C. Barrett, and R. Hughey, *Hidden Markov models for detecting remote protein homologies.* Bioinformatics, 1998. **14**(10): p. 846-56.

11. Park, J., et al., *Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods.* J Mol Biol, 1998. **284**(4): p. 1201-10.

12. Gough, J., et al., *Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.* J Mol Biol, 2001. **313**(4): p. 903-19.

13. Frishman, D., et al., *The PEDANT genome database.* Nucleic Acids Res, 2003. **31**(1): p. 207-11.

14. Buchan, D.W., et al., *Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database.* Genome Res, 2002. **12**(3): p. 503-14.

Chapter 3 Phylogeny by Protein Domain Content

3.1 Introduction

The advent of the era of complete genome sequences has led to a variety of approaches for determining the evolutionary history of organisms over and beyond the comparison of the sequences themselves [1-4], including the use of such features as concatenated protein sequences [5, 6], gene content [1-3, 7], gene order [8-10], and the distribution of structural folds [11-15]. Such efforts have continued even though there are those who feel the construction of a unified phylogeny is a hopeless task, horizontal gene transfers having been too pervasive to allow a singular depiction [16]. In this vein, it is fair to say that the resulting phylogenies have not been entirely consistent between one method and another, and certainly none on its own has resulted in a wholly satisfactory classification. Attempts to filter out anomalies [17] or the use of combinations of various approaches [9, 10] have been more satisfactory, but incongruities remain.

The principal goal of these endeavors is to generate a phylogeny that best represents the evolutionary histories of the taxa represented, and that resolves previous incongruities. It is generally agreed that three major forces are at work in modifying the genetic information in any genome: (i) expansion (gene duplication), (ii) deletion (gene loss), and (iii) exchange (horizontal transfer) [18-22]. Additionally, there must be some degree of de novo "gene genesis," the concoction of new genes by various means [23]. The challenge is to find the level of informational bundling that best accounts for this combination of events.

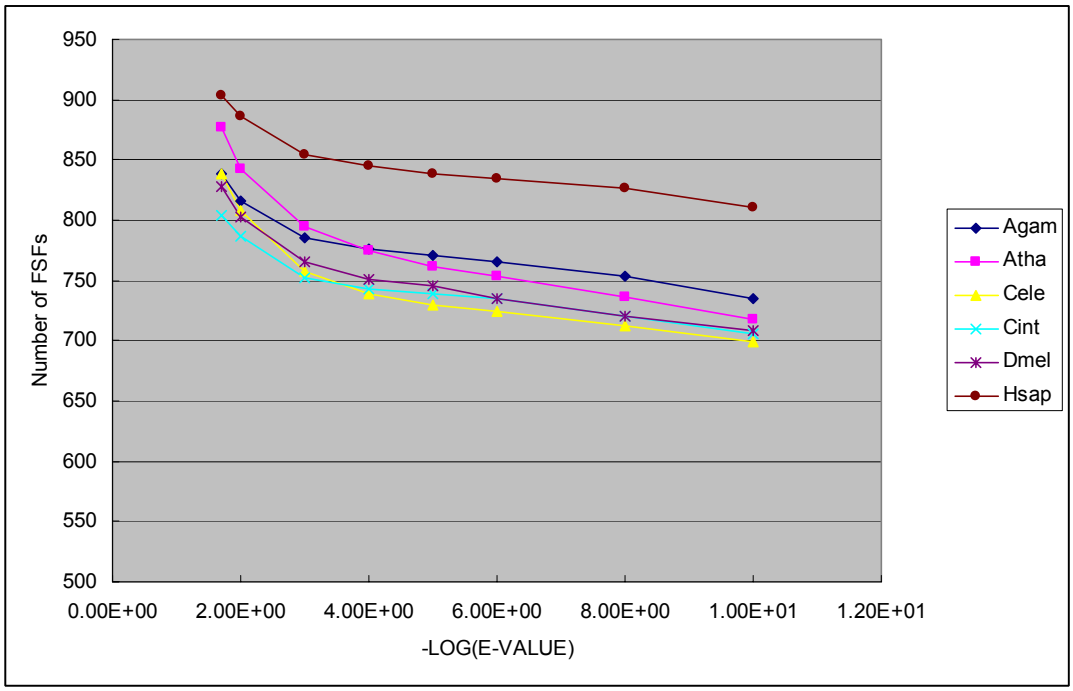
We develop a simple scheme that uses a structural attribute, the protein domain content, as the principal determinant of relatedness. In particular, we have focused on the protein domain superfamily level as opposed to the fold grouping itself that has been used by many other workers in the past [11-15]. It is a subtle but critical distinction [14]. The mere presence or absence of a domain superfamily in a genome, as opposed to its overall abundance, was used as the raw material for classification. In an examination of 174 organisms whose complete genomes have been determined, the method readily distinguishes the three major groupings of life. Beyond that, it correctly divides the Archaea into crenarchs and euryarchs and groups the Eukaryote into animals, plants, fungi, and others (protists). The most challenging part of the phylogeny reconstruction involved the 119 Bacteria, many of which are parasites and have greatly reduced genomes compared with their nearest relatives. When a weighting scheme that takes account of genome size (actually specific domain content) was used, these organisms were mostly clustered together with their proper groups.

3.2 Method

3.2.1 Data Sources

SUPERFAMILY Database [24, 25] Release 2004-07-25 covers 212 complete genomes, but many of the entries are for strains of the same species. In the study described here, only a single strain of any given species was used, reducing the number of genomes to 174. These genomes include 19 Archaea, 36 Eukarya, and 119 Bacteria. At this point ≈60% of the ORFs in the 174 completed genomes have been assigned to domain superfamilies [25]. The Superfamily database hidden Markov model searching protocol employs a probability cutoff of $E = 2 \times 10^{-2}$ for identifying likely members of a group; it also provides a confidence level (in the form of an E value) for every candidate identified. At the outset of our study, we had found that a plot of the number of alleged superfamily members versus e value for each of the major life groupings (Archaea, Bacteria, and Eukaryote) showed a sharp inflection point for each group at an E value of 10^{-4}, suggesting that a surer gauge of superfamilies needed a more stringent cutoff (Figure 3.1). Accordingly, we omitted all entries with E values greater than $E = 10^{-4}$. This cutoff reduced the total number of superfamilies assigned to any organism by 8–10%. More recent postings at the Superfamily web site have data which, when plotted as described above, show a smooth increase of the domains included, meaning no inflection point is evident. Uncertain as to what cutoff to use, we studied three sets of superfamily folds where the cutoffs were set at $E = 10^{-6}$, $E = 10^{-4}$, and $E = 2 \times 10^{-2}$, respectively, and made phylogenetic trees from them. The differences in trees were negligible, and in the end we stuck to our original cutoff point of $E = 10^{-4}$.

A



B

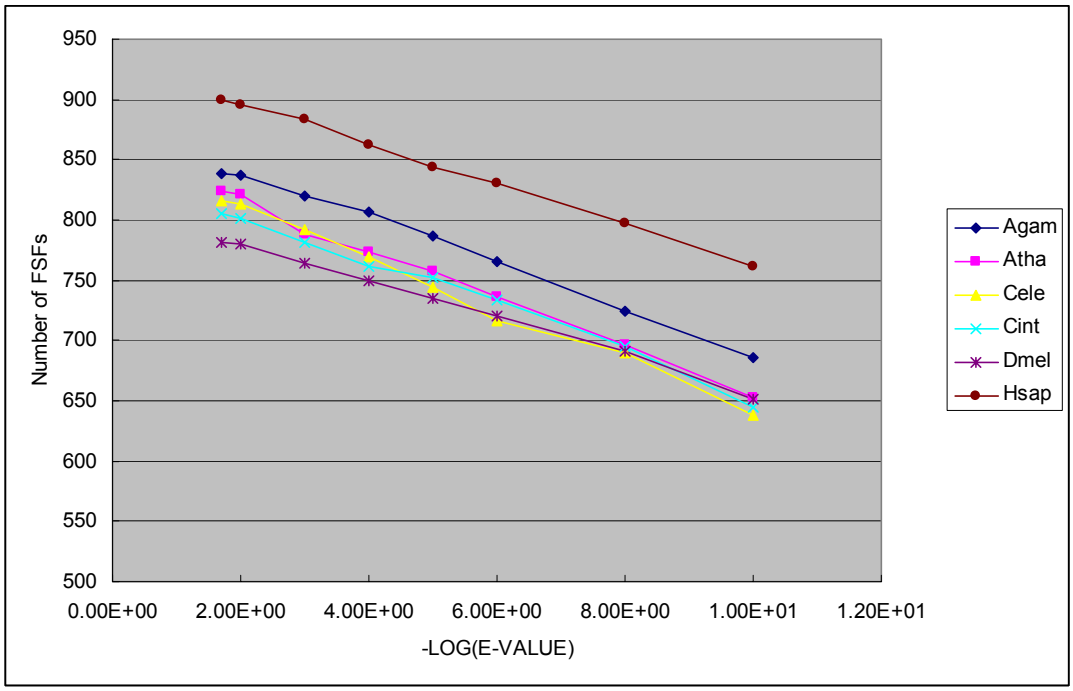


Figure 3.1 The comparison of Fold number vs. e-value cutoff figure between different release of the SUPERFAMILY database. (A) The genomic domain assignment of early version of SUPERFAMILY (Jan 06, 2004). (B) The genomic domain assignment of current version of SUPERFAMILY (July 25, 2004)

3.2.2 Data Management

The Superfamily listing of folds was downloaded and the information stored in a simple matrix with 174 columns corresponding to the organisms and 1,294 rows corresponding to the Superfamilies. The presence or absence of a particular Superfamily in a given organism was denoted with a value of 1 or 0, respectively (Table 3.1 a). Similar approaches have been used in the past [11]. Pairwise distances were calculated by considering pairs of taxa and subtracting the values in each cell from the corresponding pairmate. Nonzero values were tallied and stored in a distance matrix. A distance matrix compiled from abundances was treated similarly after normalization across all pairs.

Table 3.1 (A) Data matrix of presence/absence of protein domains in various species

	P. falciparum	S. cerevisiae	A. thaliana	C. elegans	H.sapiens	…
a.1.1	0	1	1	1	1	…
a.1.2	1	1	1	1	1	…
a.100.1	1	1	1	1	1	…
a.101.1	0	0	0	0	1	…
a.102.1	0	1	1	1	1	…
a.102.2	0	1	1	1	1	…
a.102.4	1	1	1	1	1	…
…	…	…	…	…	…	…

Each cell represents the presence/absence (denoted by 1/0) of a protein domain (in the SCOP superfamily level) in each species. Till Jan 2004, the total numbers of Superfamilies is 1234.

(B) Data matrix of the protein domain abundance in various species

	P. falciparum	S. cerevisiae	A. thaliana	C. elegans	H.sapiens	…
a.1.1	0	2	5	44	18	…
a.1.2	0	1	7	4	4	…
a.100.1	3	12	19	23	14	…
a.101.1	0	0	0	0	6	…
a.102.1	0	3	44	21	13	…
a.102.2	0	3	5	10	9	…
a.102.4	3	5	70	11	27	…
…	…	…	…	…	…	…

In contrast to Table 3.1 a, each raw value represents the numbers of a protein domain (in the SCOP superfamily level) in each species.

The great differences in genome size and gene content among the Bacteria can confound the protein domain content approach, and a weighting factor was needed that could take account of massive gene loss. In those cases, distances were corrected according to the following relationship:

$$D = A'/\left(A'+AB\right)$$

where A' is number of unique superfamily folds in the smaller of two genomes, A and B, and AB is the number of superfamily folds they share. As an example, consider the case of two closely related bacteria, one free-living and the other an intracellular parasite with a reduced genome (Figure 3.2). Clearly, the free-living species should contain most of the domains found in the parasite, and the number unique to the parasitic species (A') ought to be very small. The two tendencies are acknowledged by setting the evolutionary distance equal to the ratio of the unique domains in the smaller genome (A') to its total number of domains (A' + AB). In the limit, if the parasite has no unique domains relative to the free-living species, which is to say it had only experienced massive domain losses, the evolutionary distance would remain at zero. Similar (but different) procedures for weighting have been used by others in the past [9, 14].

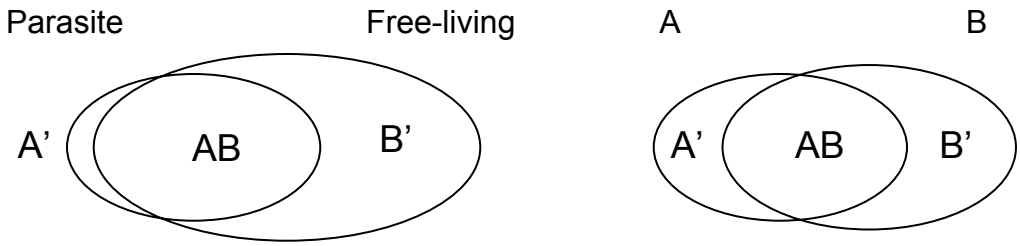


Figure 3.2 The illustration of the protein domains shared by closely related free-living and parasitic organisms.

3.2.3 Phylogenetic Methods

Phylogeny construction was performed with programs available from the PHYLIP [26] web site (http://evolution.genetics.washington.edu/phylip.html). The procedures we used were the unweighted pair group method with arithmetic mean [27] and the neighbor-joining method (NJ) [28], in each case with and without bootstrapping.

Trees were drawn with TREEVIEW (taxonomy.zoology.gla.ac.uk/rod/treeview.html). Although bootstrapping was applied to all trees, the large numbers of taxa involved made it difficult to append bootstrap values to every node in every illustration. As such, values have been appended to the figures selectively at major branch points.

In the illustrations, taxa are mostly abbreviated with an uppercase letter for genus and three lowercase letters for species (e.g., Ecol). Some exceptions to this rule were needed to avoid duplication.

3.3 Phylogenetic Trees

3.3.1 Results

Eukaryote

A simple NJ phylogenetic tree based on the presence or absence of SCOP domain Superfamilies had major clades consisting of animals, fungi, and plants, respectively, all with high bootstrap values (Figure 3.3). Within the animal clade, organisms were well positioned, the single exception being that *Xenopus tropicalis* branched off below the two fish. However, a check of the *Xenopus* superfamily listing in Superfamily revealed 74 entries not found in any other vertebrate, 40 of which were not present in any other animal, and 19 not present in any other eukaryotes. The majority of these entries were bacterial in nature, and it seemed possible that the anomalous position of *X. tropicalis* was due to contaminating sequences. Accordingly, three additional Eukaryote trees were regenerated in which the 19, 40, and 74 suspect Superfamilies were omitted, respectively. When the 19 Superfamilies not found in other eukaryotes were omitted, the branching order remained the same, but when the 40 not found in other animals were put aside, *X. tropicalis* assumed its correct position, branching off between fish and birds (chicken, Ggal). When all 74 suspect sequences were removed, *X. tropicalis* branched off between the bird and mammals. The experiment strongly suggests that the preliminary *X. tropicalis* genome sequence is contaminated. Alternative tree constructions are found in Figure 3.4.

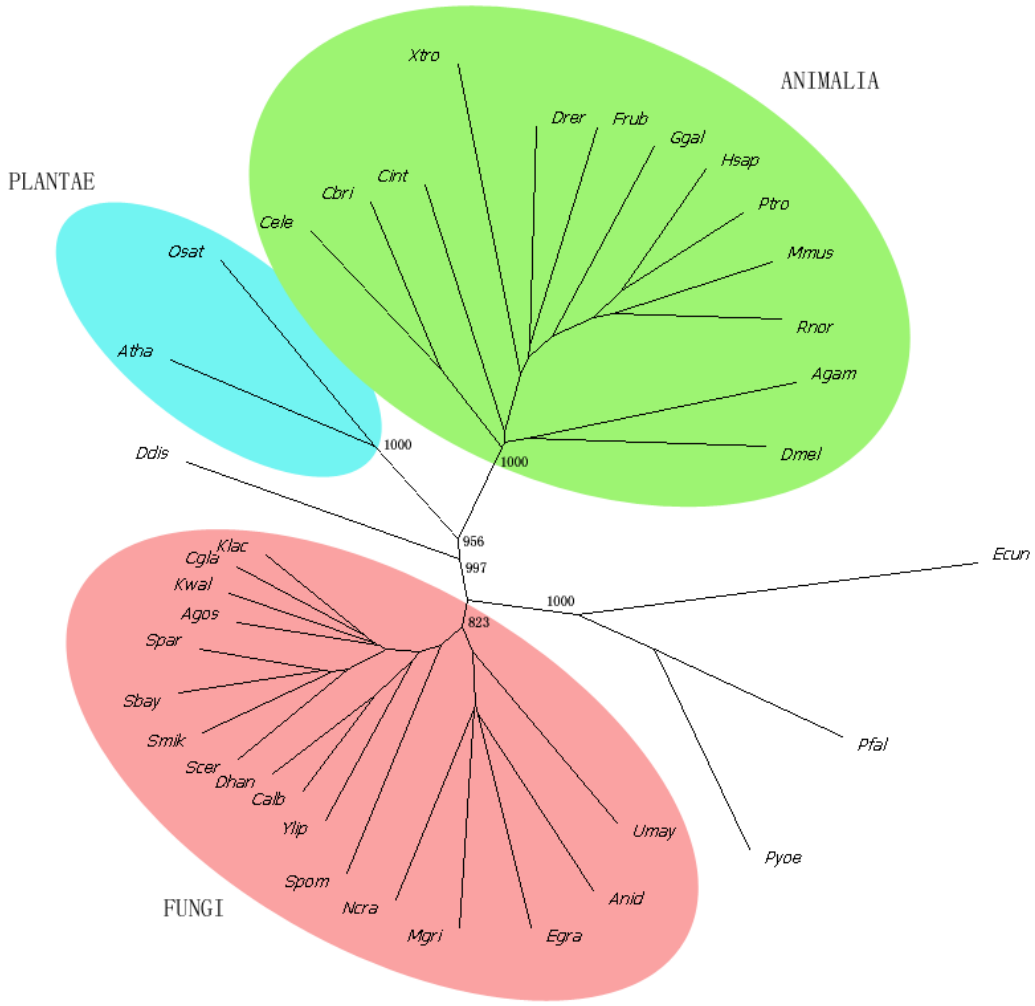


Figure 3.3 NJ tree of 36 Eukaryote based exclusively on presence or absence of SCOP superfamily folds. Bootstrap numbers are limited to the major nodes.

A

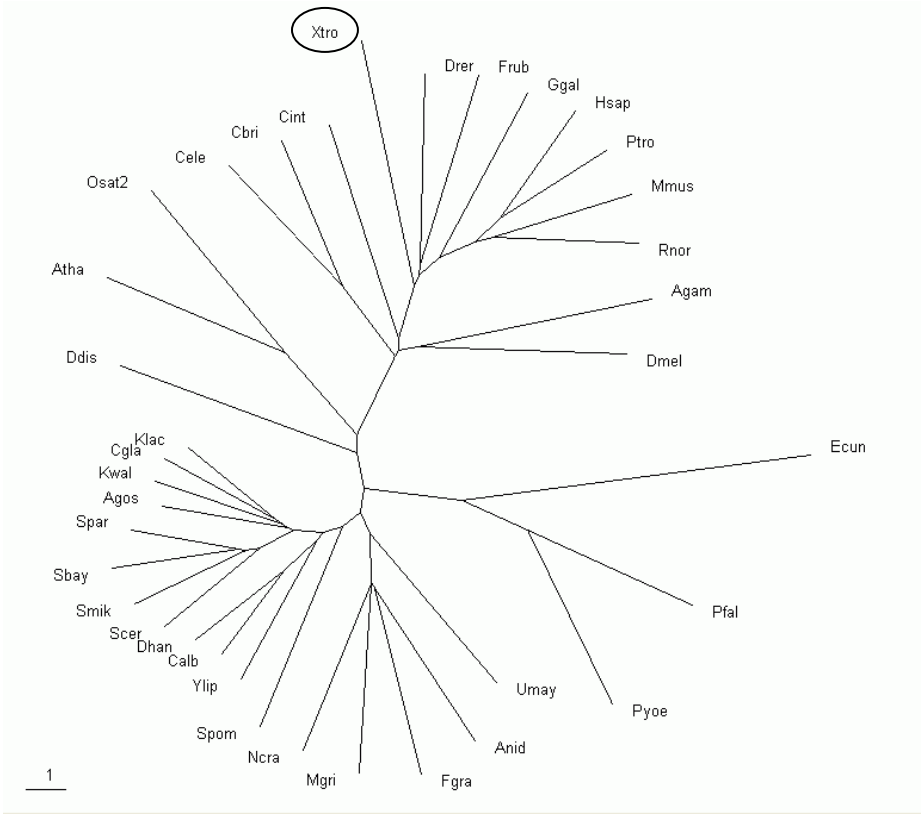
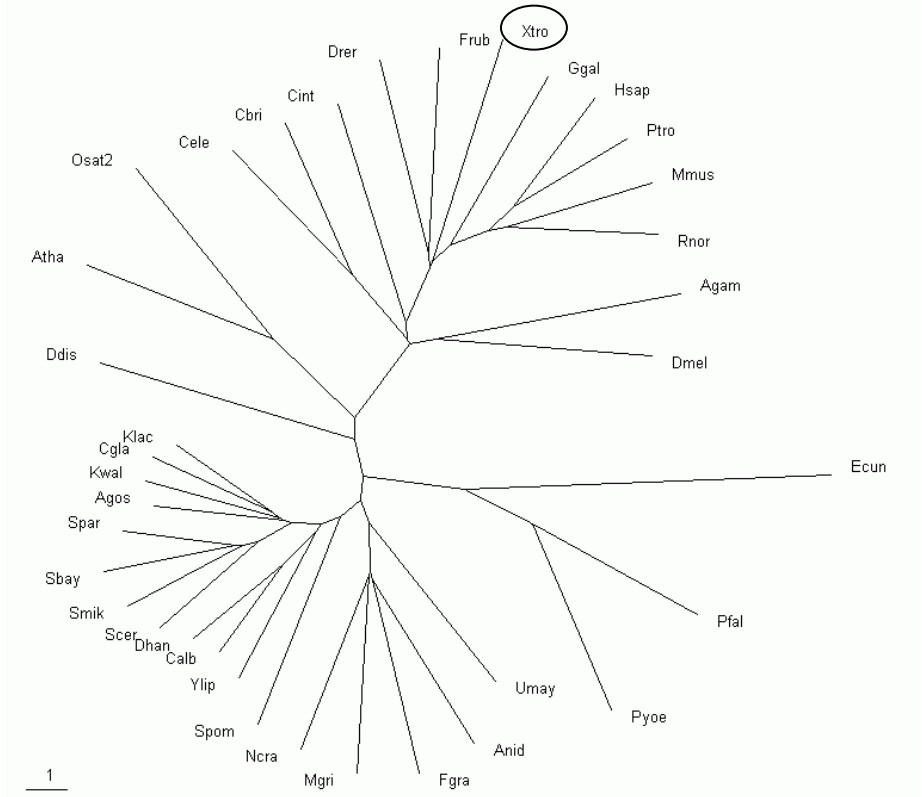


Figure 3.4 The unweighted neighbor-joining tree with the following types of protein domains removed from X. tropicalis: (A) 19 non-Eukaryote domains; (B) 40 non-animal domains; (C) 74 non-vertebrate domains.

B



C

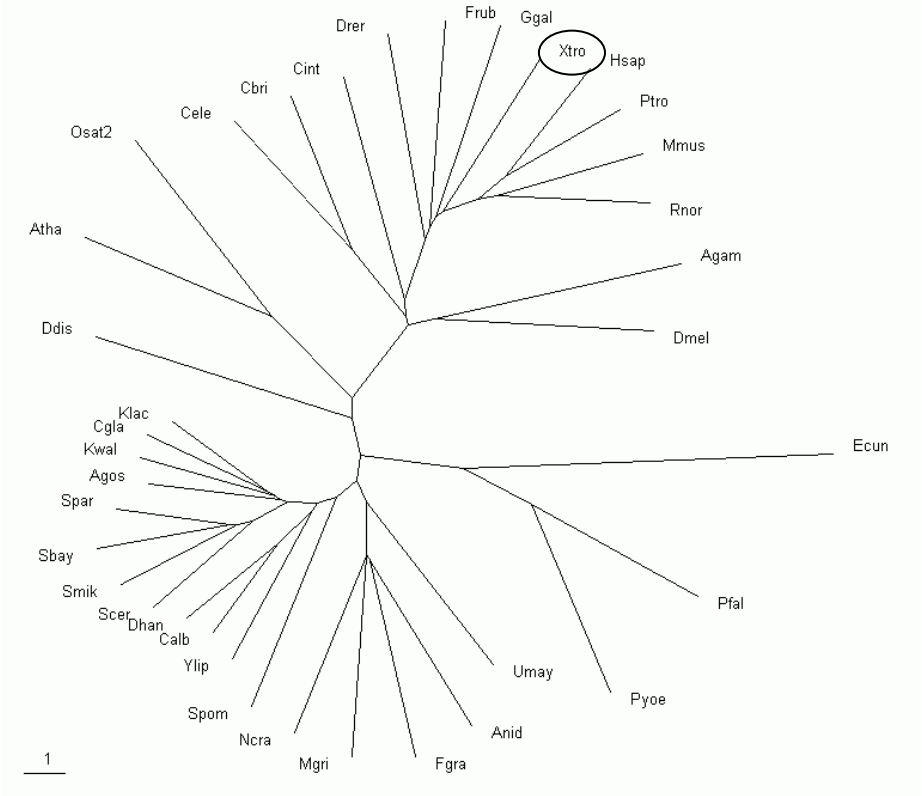


Figure 3.4 continued

Archaea

The 19 Archaea genomes were readily divided into the four Crenarchaeota and 15 Euryarchaeota (Figure 3.5). Beyond that, all of the methanogens fell into a single clade, which also included the sulfate-reducing *Archaeoglobus fulgidus*. The three *Pyrococci* bunched together (Pfur, Paby, and Phor), as did the three *Thermoplasmata* (Tvol, Taci, and Ptor). The small-genomed and enigmatic *Nanoarchaeum equitans*, reportedly the only known archaeal parasite [29], appeared near the root of the branch leading to the *Pyrococci.* When the weighting factor was used, *N. equitans* moved higher on to the same branch, and the bootstrap value improved from 0.595 to 0.968.

Halobacteria appeared on a separate basal branch, in agreement with trees based on gene content [1] and gene order [9] but not with trees based on ribosomal RNA, which generally cluster *Halobacteria* with *Methanosarcinae* [29].

All told, 13 different classical phyla were represented among the 119 bacterial genomes studied. These phyla included the five classes of 53 different proteobacteria (14 α, 7 β, 25 γ, 3 δ, and 4 ε) and four different classes of 31 Fermicutes (9 Mollicutes, 10 *Bacillales*, 8 *Lactobacillales*, and 4 *Clostridia*). Several groups included parasitic representatives with severely reduced genomes; indeed, the range of genome sizes among the Bacteria is >20-fold.

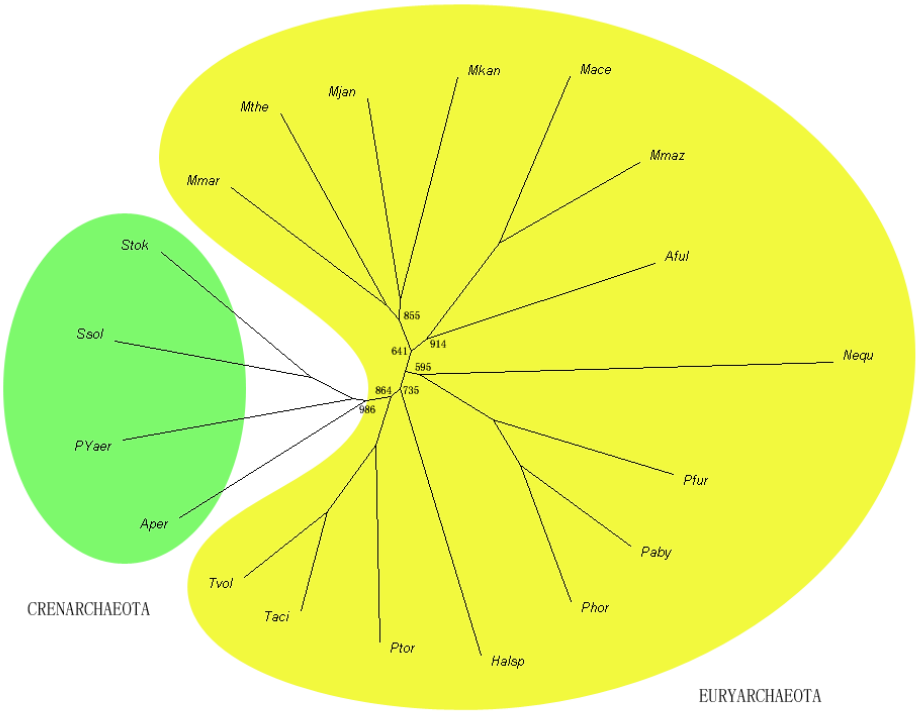


Figure 3.5 NJ tree of 19 Archaea based exclusively on presence or absence of SCOP superfamily folds. Bootstrap numbers are limited to the major nodes.

Bacteria

Phylogenies were generated with both the unweighted pair group method with arithmetic mean and NJ procedures, with and without the weighting factor, and the trees evaluated with regard to how the phyla were distributed. In fact, there were significant differences among the four sets of results. With both methods, use of the weighting factor helped provide consistency in terms of keeping members of a given phylum clustered together. The phylogenetic tree constructed by the NJ method with the weighting factor was judged the best (Figure 3.6). The comparison of weighted and unweighted UPGMA trees is in Figure 3.7.

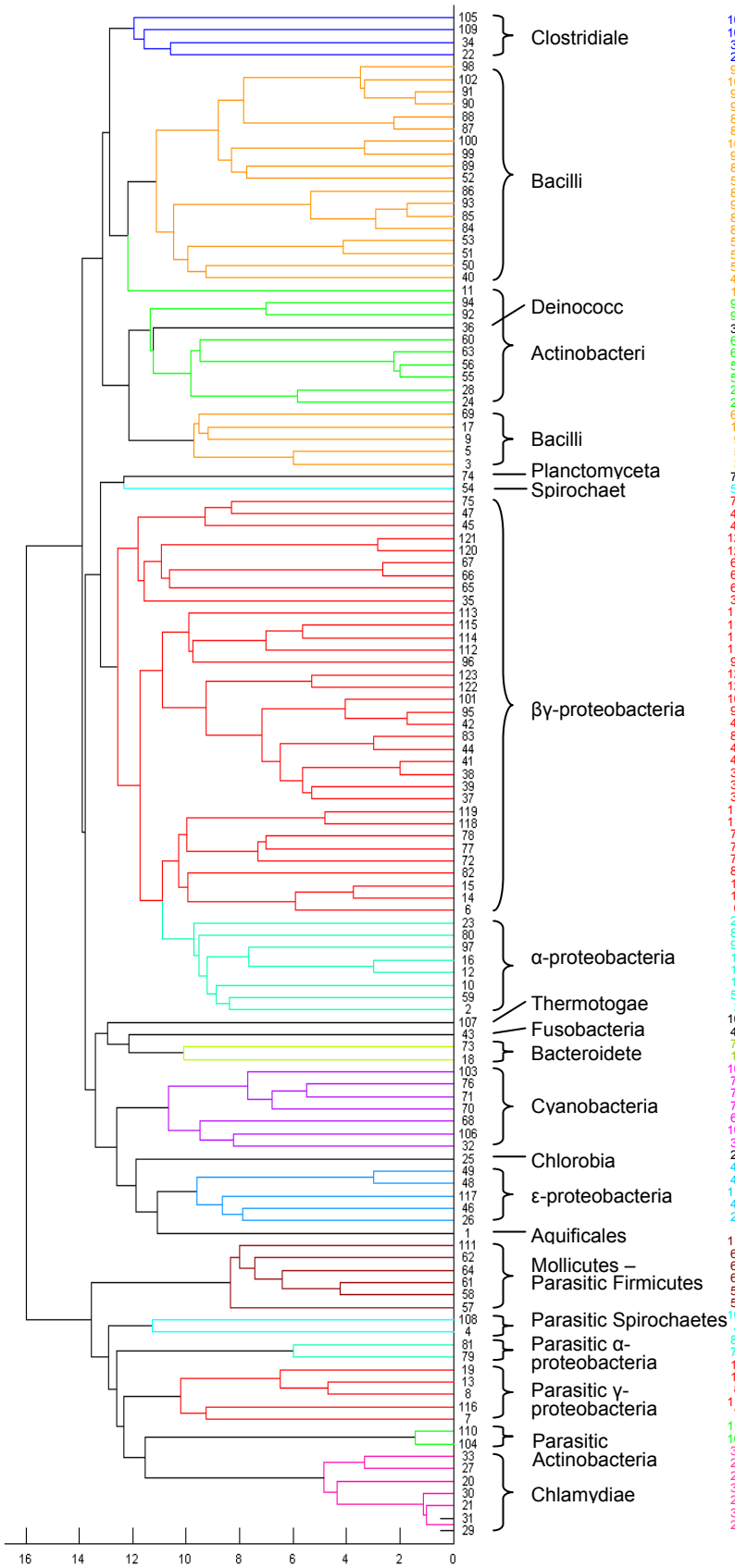
The NJ tree with weighting divided the 119 taxa into several major sectors corresponding to phyla, including one that contained all 53 proteobacteria, a subsector of the six cyanobacteria, and a branch to the lone representative of the *Aquficiae* (Figure 3.6). The two *Deinococci* appear anomalously among the proteobacteria (major anomalies are denoted with asterisks). Other major groups include all 31 fermicutes, the 11 *Actinobacteria*, and the 5 *Chlamydiae.* In a significant anomaly, the four spirochaetes separate into three sectors. In the case of the Mollicutes, eight of the nine entries cluster together, but onion yellow phytoplasma forms a separate branch. Three of four *Clostridia* cluster together, but *Thermoanaerobacter tengcongensis* (Tten) is grouped with *Thermotoga* (Tmar). The remaining phyla are discretely and satisfactorily positioned; these are the two *Bacteroidetes* and the lone taxa of *Chlorobi, Fusobacteria, Planctomycetes*, and *Thermotogae*, respectively (Figure 3.6).



Figure 3.6 NJ tree (with weighting) of 119 Bacteria based exclusively on the presence or absence of SCOP superfamily folds. Asterisks denote anomalously positioned taxa.

Figure 3.7 Unweighted (A) and weighted (B) PDC-based phylogenetic tree of 123 Bacteria organisms. All the major lineages at the order level are denoted by different color. The genome sizes of all species are listed.

A



#	Genome Size(Mbp)	Genome Name
105	2.80	Clostridium tetani E88
109	2.69	Thermoanaerobacter tengcongensis
34	3.03	Clostridium perfringens 13
22	3.94	Clostridium acetobutylicum
98	1.85	Streptococcus pyogenes M1 GAS
102	1.90	Streptococcus pyogenes MGAS8232
91	1.89	Streptococcus pyogenes SSI-1
90	1.90	Streptococcus pyogenes MGAS315
88	2.21	Streptococcus agalactiae NEM316
87	2.16	Streptococcus agalactiae 2603V/R
100	2.04	Streptococcus pneumoniae R6
99	2.16	Streptococcus pneumoniae TIGR4
89	2.03	Streptococcus mutans UA159
52	2.37	Lactococcus lactis ssp. lactis
86	2.50	Staphylococcus epidermidis ATCC 12228
93	2.88	Staphylococcus aureus ssp. aureus Mu50
85	2.81	Staphylococcus aureus ssp. aureus N315
84	2.82	Staphylococcus aureus ssp. aureus MW2
53	2.94	Listeria monocytogenes EGD-e
51	3.01	Listeria innocua
50	3.31	Lactobacillus plantarum WCFS1
40	3.22	Enterococcus faecalis V583
11	2.26	Bifidobacterium longum NCC2705
94	8.87	Streptomyces coelicolor A3(2)
92	9.03	Streptomyces avermitilis MA-4680
36	3.06	Deinococcus radiodurans
60	3.27	Mycobacterium leprae
63	4.40	Mycobacterium tuberculosis CDC1551
56	4.35	Mycobacterium bovis ssp. bovis AF2122/97
55	4.41	Mycobacterium tuberculosis H37Rv
28	3.15	Corynebacterium efficiens YS-314
24	3.31	Corynebacterium glutamicum ATCC 13032
69	3.63	Oceanobacillus iheyensis HTE831
17	4.21	Bacillus subtilis ssp. subtilis 168
9	4.20	Bacillus halodurans
5	5.41	Bacillus cereus ATCC 14579
3	5.23	Bacillus anthracis Ames
74	7.15	Pirellula sp.
54	4.63	Leptospira interrogans ser. lai 56601
75	2.26	Pasteurella multocida
47	1.83	Haemophilus influenzae Rd
45	1.70	Haemophilus ducreyi 35000HP
121	2.52	Xylella fastidiosa Temecula1
120	2.68	Xylella fastidiosa 9a5c
67	2.18	Neisseria meningitidis Z2491
66	2.27	Neisseria meningitidis MC58
65	2.81	Nitrosomonas europaea ATCC 19718
35	2.00	Coxiella burnetii RSA 493
113	4.75	Chromobacterium violaceum ATCC 12472
115	5.12	Vibrio vulnificus CMCP6
114	5.16	Vibrio parahaemolyticus RIMD 2210633
112	5.03	Vibrio cholerae
96	4.97	Shewanella oneidensis MR-1
123	4.65	Yersinia pestis CO92
122	4.60	Yersinia pestis KIM
101	4.86	Salmonella typhimurium LT2
95	4.79	Salmonella enterica ssp. enterica ser. Typhi
42	4.81	Salmonella enterica ssp. enterica ser. Typhi Ty2
83	4.61	Shigella flexneri 2a 301
44	4.60	Shigella flexneri 2a 2457T
41	5.50	Escherichia coli O157:H7
38	5.53	Escherichia coli O157:H7 EDL933
39	4.64	Escherichia coli K12
37	5.23	Escherichia coli CFT073
119	5.08	Xanthomonas campestris pv. campestris ATCC 33913
118	5.18	Xanthomonas axonopodis pv. citri 306
78	6.40	Pseudomonas syringae pv. tomato DC3000
77	6.18	Pseudomonas putida KT2440
72	6.26	Pseudomonas aeruginosa PAO1
82	3.72	Ralstonia solanacearum
15	4.77	Bordetella parapertussis
14	5.34	Bordetella bronchiseptica
6	4.09	Bordetella pertussis
23	4.02	Caulobacter crescentus CB15
80		Rhodobacter sphaeroides
97	3.65	Sinorhizobium meliloti
16	3.31	Brucella suis 1330
12	3.29	Brucella melitensis 16M
10	9.12	Bradyrhizobium japonicum USDA 110
59	7.04	Mesorhizobium loti
2	9.83	Agrobacterium tumefaciens C58
107	1.86	Thermotoga maritima
43	2.17	Fusobacterium nucleatum ssp. nucleatum ATCC 25586
73	2.34	Porphyromonas gingivalis W83
18	6.26	Bacteroides thetaiotaomicron VPI-5482
103	2.43	Synechococcus sp. WH 8102
76	1.75	Prochlorococcus marinus ssp. marinus CCMP1375
71	1.66	Prochlorococcus marinus ssp. pastoris CCMP1378
70	2.41	Prochlorococcus marinus MIT 9313
68	6.41	Nostoc sp. PCC 7120
106	2.59	Thermosynechococcus elongatus BP-1
32	3.57	Synechocystis sp. PCC 6803
25	2.15	Chlorobium tepidum TLS
49	1.64	Helicobacter pylori J99
48	1.68	Helicobacter pylori 26695
117	2.11	Wolinella succinogenes
46	1.80	Helicobacter hepaticus ATCC 51449
26	1.64	Campylobacter jejuni ssp. jejuni NCTC 11168
1	1.55	Aquifex aeolicus VF5
111	0.75	Ureaplasma urealyticum
62	0.96	Mycoplasma pulmonis
64	1.00	Mycoplasma gallisepticum R
61	0.82	Mycoplasma pneumoniae
58	0.58	Mycoplasma genitalium
57	1.36	Mycoplasma penetrans
108	1.14	Treponema pallidum
4	0.91	Borrelia burgdorferi B31
81	1.11	Rickettsia prowazekii
79	1.23	Rickettsia conorii
19	0.62	Buchnera aphidicola Bp
13	0.64	Buchnera aphidicola APS
8	0.64	Buchnera aphidicola Sg
116	0.70	Wigglesworthia glossinidia
7	0.71	Candidatus Blochmannia floridanus
110	0.93	Tropheryma whipplei Twist
104	0.93	Tropheryma whipplei TW08/27
33	1.04	Chlamydia trachomatis
27	1.07	Chlamydia muridarum
20	1.17	Chlamydophila caviae GPIC
30	1.23	Chlamydophila pneumoniae AR39
21	1.23	Chlamydophila pneumoniae TW-183
31	1.23	Chlamydophila pneumoniae CWL029
29	1.23	Chlamydophila pneumoniae J138



#	Genome Size(Mbp)	Genome Name
107	1.86	Thermotoga maritima
108	1.14	Treponema pallidum
4	0.91	Borrelia burgdorferi B31
43	2.17	Fusobacterium nucleatum ssp. nucleatum ATCC 25586
109	2.69	Thermoanaerobacter tengcongensis
105	2.8	Clostridium tetani E88
34	3.03	Clostridium perfringens 13
22	3.94	Clostridium acetobutylicum
100	2.04	Streptococcus pneumoniae R6
99	2.16	Streptococcus pneumoniae TIGR4
102	1.9	Streptococcus pyogenes MGAS8232
98	1.85	Streptococcus pyogenes M1 GAS
91	1.89	Streptococcus pyogenes SSI-1
90	1.9	Streptococcus pyogenes MGAS315
88	2.21	Streptococcus agalactiae NEM316
87	2.16	Streptococcus agalactiae 2603V/R
89	2.03	Streptococcus mutans UA159
52	2.37	Lactococcus lactis ssp. lactis
53	2.94	Listeria monocytogenes EGD-e
51	3.01	Listeria innocua
50	3.31	Lactobacillus plantarum WCFS1
40	3.22	Enterococcus faecalis V583
9	4.2	Bacillus halodurans
86	2.5	Staphylococcus epidermidis ATCC 12228
93	2.88	Staphylococcus aureus ssp. aureus Mu50
85	2.81	Staphylococcus aureus ssp. aureus N315
84	2.82	Staphylococcus aureus ssp. aureus MW2
69	3.63	Oceanobacillus iheyensis HTE831
17	4.21	Bacillus subtilis ssp. subtilis 168
5	5.41	Bacillus cereus ATCC 14579
3	5.23	Bacillus anthracis Ames
111	0.75	Ureaplasma urealyticum
62	0.96	Mycoplasma pulmonis
64	1	Mycoplasma gallisepticum R
61	0.82	Mycoplasma pneumoniae
58	0.58	Mycoplasma genitalium
57	1.36	Mycoplasma penetrans
73	2.34	Porphyromonas gingivalis W83
18	6.26	Bacteroides thetaiotaomicron VPI-5482
106	2.59	Thermosynechococcus elongatus BP-1
68	6.41	Nostoc sp. PCC 7120
103	2.43	Synechococcus sp. WH 8102
71	1.66	Prochlorococcus marinus ssp. pastoris CCMP1378
76	1.75	Prochlorococcus marinus ssp. marinus CCMP1375
70	2.41	Prochlorococcus marinus MIT 9313
32	3.57	Synechocystis sp. PCC 6803
36	3.06	Deinococcus radiodurans
60	3.27	Mycobacterium leprae
63	4.4	Mycobacterium tuberculosis CDC1551
56	4.35	Mycobacterium bovis ssp. bovis AF2122/97
55	4.41	Mycobacterium tuberculosis H37Rv
110	0.93	Tropheryma whipplei Twist
104	0.93	Tropheryma whipplei TW08/27
94	8.87	Streptomyces coelicolor A3(2)
92	9.03	Streptomyces avermitilis MA-4680
28	3.15	Corynebacterium efficiens YS-314
24	3.31	Corynebacterium glutamicum ATCC 13032
11	2.26	Bifidobacterium longum NCC2705
49	1.64	Helicobacter pylori J99
48	1.68	Helicobacter pylori 26695
46	1.8	Helicobacter hepaticus ATCC 51449
117	2.11	Wolinella succinogenes
26	1.64	Campylobacter jejuni ssp. jejuni NCTC 11168
35	2	Coxiella burnetii RSA 493
96	4.97	Shewanella oneidensis MR-1
115	5.12	Vibrio vulnificus CMCP6
114	5.16	Vibrio parahaemolyticus RIMD 2210633
112	5.03	Vibrio cholerae
47	1.83	Haemophilus influenzae Rd
75	2.26	Pasteurella multocida
45	1.7	Haemophilus ducreyi 35000HP
123	4.65	Yersinia pestis CO92
122	4.6	Yersinia pestis KIM
101	4.86	Salmonella typhimurium LT2
95	4.79	Salmonella enterica ssp. enterica ser. Typhi
42	4.81	Salmonella enterica ssp. enterica ser. Typhi Ty2
39	4.64	Escherichia coli K12
116	0.7	Wigglesworthia glossinidia
41	5.5	Escherichia coli O157:H7
38	5.53	Escherichia coli O157:H7 EDL933
83	4.61	Shigella flexneri 2a 301
44	4.6	Shigella flexneri 2a 2457T
37	5.23	Escherichia coli CFT073
19	0.62	Buchnera aphidicola Bp
13	0.64	Buchnera aphidicola APS
8	0.64	Buchnera aphidicola Sg
121	2.52	Xylella fastidiosa Temecula1
120	2.68	Xylella fastidiosa 9a5c
119	5.08	Xanthomonas campestris pv. campestris ATCC 33913
118	5.18	Xanthomonas axonopodis pv. citri 306
65	2.81	Nitrosomonas europaea ATCC 19718
113	4.75	Chromobacterium violaceum ATCC 12472
67	2.18	Neisseria meningitidis Z2491
66	2.27	Neisseria meningitidis MC58
82	3.72	Ralstonia solanacearum
77	6.18	Pseudomonas putida KT2440
78	6.4	Pseudomonas syringae pv. tomato DC3000
72	6.26	Pseudomonas aeruginosa PAO1
7	0.71	Candidatus Blochmannia floridanus
15	4.77	Bordetella parapertussis
14	5.34	Bordetella bronchiseptica
6	4.09	Bordetella pertussis
81	1.11	Rickettsia prowazekii
79	1.23	Rickettsia conorii
23	4.02	Caulobacter crescentus CB15
80		Rhodobacter sphaeroides
16	3.31	Brucella suis 1330
12	3.29	Brucella melitensis 16M
10	9.12	Bradyrhizobium japonicum USDA 110
97	3.65	Sinorhizobium meliloti
59	7.04	Mesorhizobium loti
2	9.83	Agrobacterium tumefaciens C58
1	1.55	Aquifex aeolicus VF5
54	4.63	Leptospira interrogans ser. lai 56601
74	7.15	Pirellula sp.
25	2.15	Chlorobium tepidum TLS
33	1.04	Chlamydia trachomatis
27	1.07	Chlamydia muridarum
20	1.17	Chlamydophila caviae GPIC
30	1.23	Chlamydophila pneumoniae AR39
31	1.23	Chlamydophila pneumoniae CWL029
29	1.23	Chlamydophila pneumoniae J138
21	1.23	Chlamydophila pneumoniae TW-183

Figure 3.7 continued

Overall Phylogeny

Both unweighted pair group method with arithmetic mean and NJ were used to construct trees for all 174 taxa. The unweighted pair group method with arithmetic mean tree readily divided the taxa into the three superkingdoms, but many bacteria with small genomes fell into a clade separate from their nearest relatives within the bacterial zone. The use of the weighting factor led to most of them being redistributed to their proper positions. In the case of the NJ tree, the three superkingdoms were mostly correct, but two organisms with reduced genomes, the archaeon *N. equitans* and the eukaryote *E. cuniculi*, fell among the Bacteria. When the weighting factor was used, the two anomalously placed small-genome organisms retreated to their correct realms, and the tripartite grouping was maintained with high bootstrap values (Figure 3.8). Concomitantly, however, a number of rearrangements occurred within the three major groups; in the end, the best phylogenetic trees were generated when the taxa were restricted to a single superkingdom.

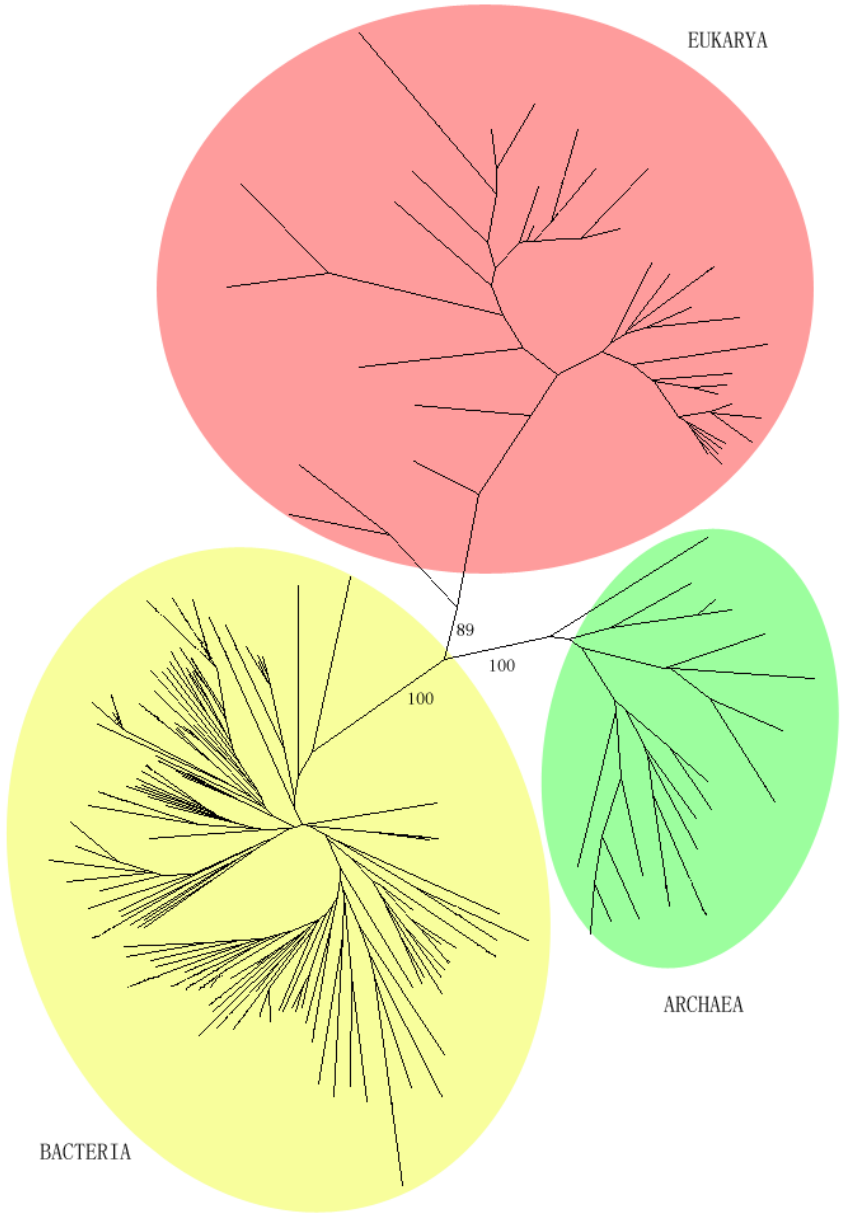


Figure 3.8 Overall phylogeny (NJ) of 174 organisms for which complete genomes have been determined. Bootstrap number was limited to the major branch point.

3.3.2 Discussion

Fold Level vs. Superfamily Level

A number of previous reports have used the fold level of classification for taxonomic purposes [11-15], and it may be useful to consider the subtle differences between the "fold level" (800 listed in SCOP [30, 31] release 1.65) and the "superfamily level" (1,294 listed). As the numbers imply, many Superfamilies have only one member, in which case the superfamily is the fold. In contrast, other fold-level categories embrace several superfamilies. By definition, the fold level of the SCOP hierarchy lists entries that have the same secondary structure and chain topology but for which there is little or no evidence of common ancestry [25]. The superfamily level is defined as comprising those folds for which there is structural and functional evidence of common ancestry, and the next level of clustering, the family, is reserved for members that have obvious sequence resemblance.

The obvious advantage of using the superfamily level is that it offers a higher level of certainty that the members of each group do in fact share common ancestry; it also provides a finer grid for classification purposes. Although many superfamily entries share common ancestry with other superfamilies assigned to the same fold, the consequences for phylogenetic reconstruction in those cases where common ancestry is not the case can be dire. In this regard, Lin and Gerstein [14] found that the superfamily level performed more ably than the fold level when the number of genomes under study was increased from eight to 20.

Presence/Absence vs. Abundance

Remarkably, the mere presence or absence of protein folds (at the superfamily level) in genomes more accurately reconstructs most of the phylogenies examined here than when the overall abundance of each domain superfamily fold in a genome was used (Table 3.1). This result was true even with regard to the threefold distribution of Archaea, Bacteria, and Eukarya because some taxa went astray when abundance was used (Figure 3.9).

Domain abundance is greatly affected by gene and chromosome duplication, which although contributing to the evolutionary distance between genomes is not a uniform process. It has long been recognized that genetic duplication begets more duplication, the natural consequence of more opportunities for homologous recombination. As a result, excessive duplication can lead to inflated distances that mask the more crucial differences in the form of gain or loss of individual domains. The protein domain content of a given genome is changed whenever (*i*) a new fold evolves during a long-term divergence, (*ii*) a fold is lost as a result of deletion of all or part of a gene, or (*iii*) a new fold is acquired by horizontal transfer. Ordinarily, genetic duplication on its own does not give rise abruptly to new folds.

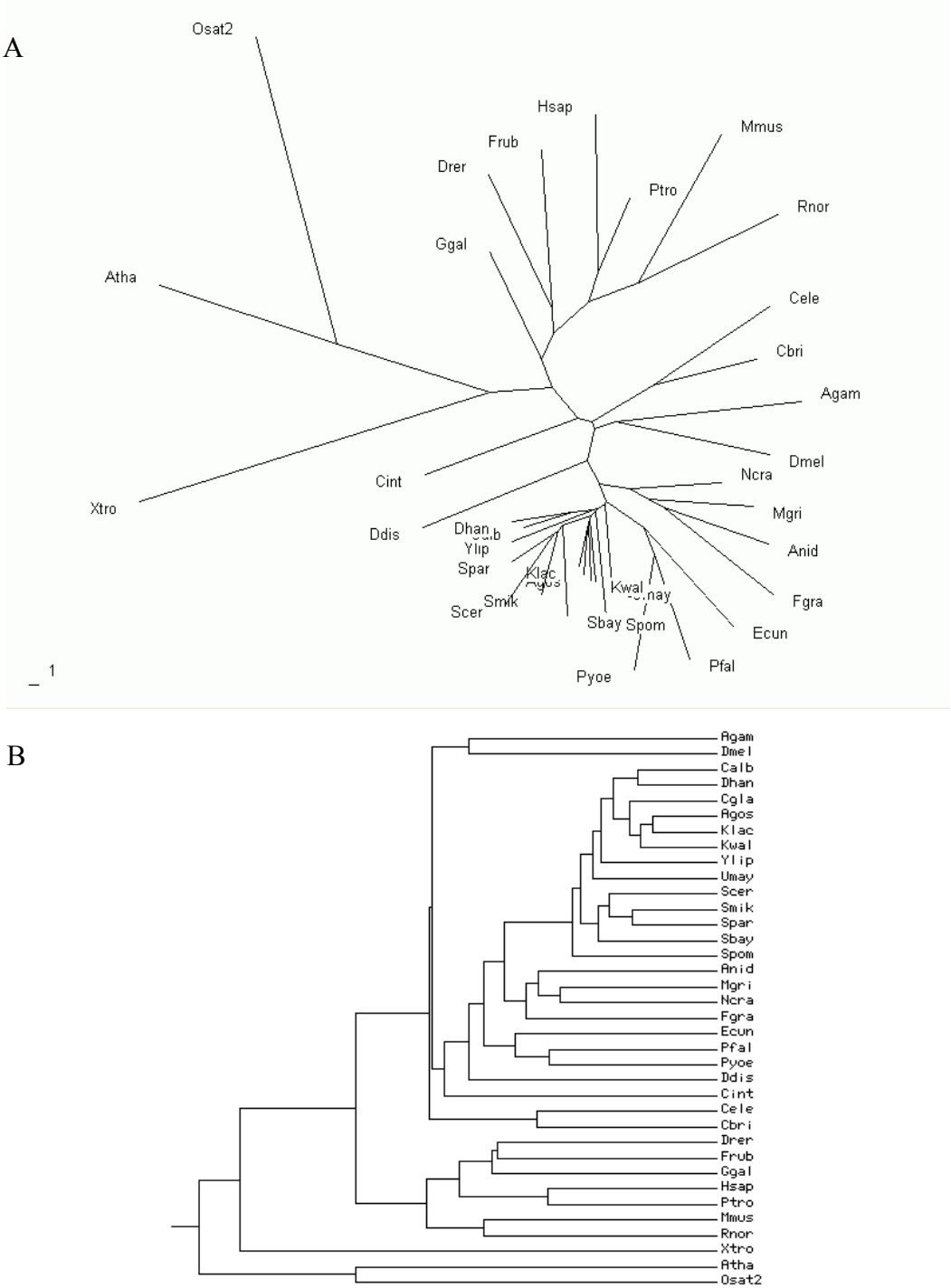


Figure 3.9 The NJ (A) and UPGMA (B) Eukaryote trees based on protein domain abundance. Plant species that have large genome size and thus more protein domains protrude from the rest of the organism in the tree.

Fold Content vs. Gene Content

It is fair to ask why the protein domain content should be any better than gene content in classifying genomes. One answer is that proteins (gene products) are modular, and many of them are mosaics of different domains [32]. Indeed, the duplicated and/or shuffled domain is the foundational unit from which new protein equipment is fashioned. Genes may be retained even when the domain content changes and vice versa. Certainly protein domain content measures evolutionary change differently from gene content.

There are some other intrinsic advantages to using the simple presence or absence of a structural attribute for phylogenetic purposes. For one, there is less concern about mistaken paralogy, as so often occurs when comparing protein sequences. Moreover, the rate of sequence change and its attendant problems of site-specific variation do not play a role, and arbitrary decisions about gene designation and function are not issues. As a general rule, also, three-dimensional structures are more highly conserved than primary sequences, allowing one to see further into the evolutionary past. It is noteworthy that in the present study, all of the decisions, arbitrary or not, about what constitutes a particular structural element and its presence or absence in a genome have been made by others (i.e., SCOP and Superfamily). As such, the results are completely objective and should be easily reproduced by anyone.

Phylogenomics

As more complete genomes are sequenced, phylogenetic analysis is entering a new era – phylogenomics. At this stage, not only one or several genes or morphological characteristics of species are considered, but the entire genomes of organisms are used to reconstruct the evolutionary history of life. This reconstruction involves the identification of homologous characters that are shared between different organisms and inference of phylogenetic trees from the analysis of these characters using reconstruction methods. The accuracy of the inference is therefore heavily dependent on both the understanding of the evolutionary model of these homologous characters and the reconstruction methods that reflect their evolutionary processes.

Provided with complete genome data, phylogenomic reconstruction methods are no longer hampered by the sampling effect due to the limited number of available characters. According to the homologous characters in comparison, phylogenomics methods can be divided in to sequence-based methods that rely on the sequence comparison at genome scale, and methods based on whole-genome features that includes gene content, gene order and protein domain content we developed. Three standard phylogenetic reconstruction methods, distance methods, maximum parsimony and maximum likelihood, are still applicable in phylogenomics.

Progresses in phylogenomics have demonstrated its strength in resolving several fundamental evolutionary problems. However, errors do occur. In sequence-based methods, variations in evolutionary rate among species can cause the long-branch-attraction artifact. In the methods based on gene content or protein domain content, the

substantial gene losses in parasitic species cause the erroneous grouping of organisms, a phenomenon called the “big genome attraction” artifact. Horizontal gene transfer (HGT) certainly plays a key role in the evolution of bacterial genomes; its influence on phylogenomic reconstruction methods has not yet been fully investigated.

Reliable phylogenomic reconstruction depends on the identification of the proper homologous characters that reflect major evolutionary processes. The ideal homologous characters are those that underwent only a few changes over time in order to prevent noise by convergence and reversal. As discussed earlier, protein domains are evolutionarily more conserved than protein sequence, in addition, their invention, loss and transfer are major evolutionary processes that only occurred occasionally. Therefore, protein domain content is a proper character for phylogenomic reconstruction, as demonstrated before.

However, because of the complicated nature of the evolutionary process that involves changes at various levels, such as substitution of nucleic acids in DNA and amino acids in proteins, domain swapping, chromosome inversion, morphological changes in phenotype, no single character can represent the entire evolutionary change. In consequence, no phylogenomic method is perfect. The reconstruction of the tree of life requires the comprehensive understanding of the evolutionary history at all level and the combination of multiple approaches based on different homologous character.

3.4 Distribution of Protein Domains in the Three Superkingdoms

3.4.1 Results

Fold Distribution

The existence and distribution pattern of all protein domains across the superkingdoms of life provides a course-grained view of their evolutionary development. We retrieved the domain assignment data from SUPERFAMILY release in 07-25-2004, which contains 174 complete genomes, including 19 Archaea, 119 Bacteria and 36 Eukaryote. (Chapter 2 includes the list of all organisms used in the study.) The Venn diagram in Figure 3.10 A shows the census of the protein domains at the SCOP Superfamily level among the three superkingdoms.

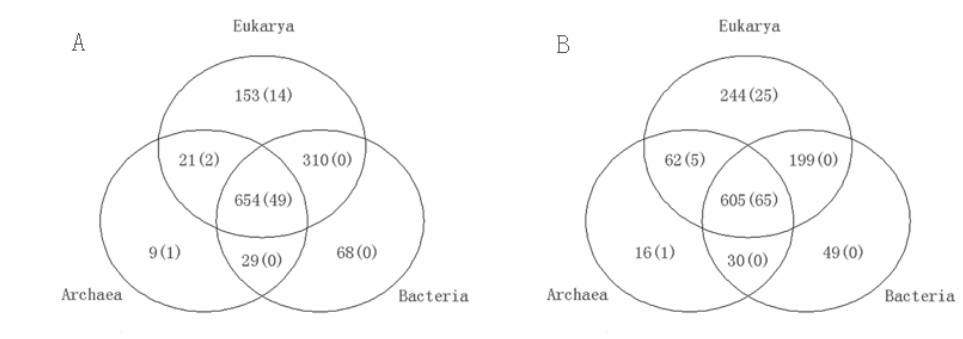


Figure 3.10 Venn diagrams showing occurrence of 1,244 Domain Superfamilies in the three superkingdoms. The first value in any sector reflects the occurrence of a particular Domain Superfamilies in any genome in that sector; values in parentheses indicate numbers found in all genomes in a sector. (A) Census based on 174 complete genomes. (B) Census based on complete genomes of 19 Archaea, 19 selected Eukarya, and 19 selected Bacteria.

Of the 1,294 Domain Superfamilies examined, 50, mostly found in viruses, did not occur in any of the 174 genomes under study. In line with previous reports [13, 15], a fold is counted if it occurs even once in a given realm. The numbers follow a similar pattern to those earlier reports [13, 15], about half of all folds occurring at some frequency in all three superkingdoms. Eukaryotes have the largest number of folds not found in the other two superkingdoms and Archaea the fewest. Bacteria share more folds with eukaryotes than they do with Archaea. Showing that these trends are more or less independent of the different numbers of taxa representing the three groups, a second computation was conducted by counting the smaller sets of genomes, beginning with one that included all 19 Archaea and 19 selected genomes from each of the other two realms (Figure 3.10 B). In subsequent counts, the number of taxa included for the Bacteria and Eukaryote was progressively increased until the full set was reached (Tables 2.3-4).

As an added, and arguably a more interesting feature, we noted the counts for those folds that are both unique and ubiquitous for a given sector (Figure 3.10 A). Thus, 49 folds occur in all 174 genomes, and 14 folds are found in all 36 eukaryotic genomes but not elsewhere. Among the bacteria, there are no unique folds that occur in all bacterial genomes but nowhere else or even that they share uniquely with only Archaea or only Eukaryotes. This finding was true even when only 19 bacterial genomes were used in the count with 19 each of the other superkingdoms (Figure 3.10 B). Interestingly, if the count is made at the fold level, the number of folds shared by all 174 genomes is 52, hardly changed from the 49 found for the Superfamily level (Tables 3.2).

Table 3.2 (A) Unique Superfamilies exist exclusively within each region: A (1), E (14), AE (2), ABE (49)

Count	Region	FSF	A	B	E	SCOP definition
1	A	d.17.6	19	0	0	Archaeosine tRNA-guanine transglycosylase, C2 domain
1	E	a.24.15	0	0	36	FAD-dependent thiol oxidase
2	E	a.29.2	0	0	36	Bromodomain
3	E	a.40.1	0	0	36	Calponin-homology domain, CH-domain
4	E	a.69.2	0	0	36	Ypt/Rab-GAP domain of gyp1p
5	E	b.34.10	0	0	36	Cap-Gly domain
6	E	b.72.1	0	0	36	WW domain
7	E	c.52.3	0	0	36	Eukaryotic RPB5 N-terminal domain
8	E	d.109.1	0	0	36	Actin depolymerizing proteins
9	E	d.148.1	0	0	36	Ubiquitin-protein ligase E3a, Hect catalytic domain (E6ap)
10	E	d.189.1	0	0	36	PX domain
11	E	d.20.1	0	0	36	UBC-like
12	E	e.40.1	0	0	36	Cullin homology domain
13	E	g.41.10	0	0	36	Zn-finger domain of Sec23/24
14	E	g.45.1	0	0	36	Pyk2-associated protein beta ARF-GAP domain
1	AE	d.230.1	19	0	36	N-terminal, heterodimerisation domain of RBP4 (RpoE)
2	AE	g.59.1	19	0	36	Zinc-binding domain of translation initiation factor 2 beta
1	ABE	a.156.1	19	119	36	S13-like H2TH domain
2	ABE	a.16.1	19	119	36	S15/NS1 RNA-binding domain
3	ABE	a.27.1	19	119	36	Anticodon-binding domain of a subclass of class I aminoacyl-tRNA synthetases
4	ABE	a.4.7	19	119	36	Ribosomal protein L11, C-terminal domain
5	ABE	a.60.7	19	119	36	5' to 3' exonuclease, C-terminal subdomain
6	ABE	a.75.1	19	119	36	Ribosomal protein S7
7	ABE	b.34.5	19	119	36	Translation proteins SH3-like domain
8	ABE	b.39.1	19	119	36	Ribosomal protein L14
9	ABE	b.40.4	19	119	36	Nucleic acid-binding proteins
10	ABE	b.43.3	19	119	36	Translation proteins
11	ABE	b.44.1	19	119	36	EF-Tu/eEF-1alpha/eIF2-gamma C-terminal domain
12	ABE	b.51.1	19	119	36	ValRS/IleRS/LeuRS editing domain
13	ABE	c.1.9	19	119	36	Metallo-dependent hydrolases
14	ABE	c.108.1	19	119	36	HAD-like
15	ABE	c.12.1	19	119	36	Ribosomal proteins L15p and L18e
16	ABE	c.120.1	19	119	36	PIN domain-like
17	ABE	c.2.1	19	119	36	NAD(P)-binding Rossmann-fold domains
18	ABE	c.20.1	19	119	36	Initiation factor IF2/eIF5b, domain 3
19	ABE	c.23.15	19	119	36	Ribosomal protein S2
20	ABE	c.23.16	19	119	36	Class I glutamine amidotransferase-like
21	ABE	c.26.1	19	119	36	Nucleotidylyl transferase
22	ABE	c.26.2	19	119	36	Adenine nucleotide alpha hydrolases-like

Table 3.2 (A) continued

Count	Region	FSF	A	B	E	SCOP definition
23	ABE	c.3.1	19	119	36	FAD/NAD(P)-binding domain
24	ABE	c.37.1	19	119	36	P-loop containing nucleoside triphosphate hydrolases
25	ABE	c.51.1	19	119	36	Anticodon-binding domain of Class II aaRS
26	ABE	c.55.1	19	119	36	Actin-like ATPase domain
27	ABE	c.55.3	19	119	36	Ribonuclease H-like
28	ABE	c.55.4	19	119	36	Translational machinery components
29	ABE	c.66.1	19	119	36	S-adenosyl-L-methionine-dependent methyltransferases
30	ABE	d.104.1	19	119	36	Class II aaRS and biotin synthetases
31	ABE	d.122.1	19	119	36	ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase
32	ABE	d.127.1	19	119	36	Creatinase/aminopeptidase
33	ABE	d.131.1	19	119	36	DNA clamp
34	ABE	d.14.1	19	119	36	Ribosomal protein S5 domain 2-like
35	ABE	d.140.1	19	119	36	Ribosomal protein S8
36	ABE	d.141.1	19	119	36	Ribosomal protein L6
37	ABE	d.142.2	19	119	36	DNA ligase/mRNA capping enzyme, catalytic domain
38	ABE	d.157.1	19	119	36	Metallo-hydrolase/oxidoreductase
39	ABE	d.26.1	19	119	36	FKBP-like
40	ABE	d.41.4	19	119	36	Ribosomal protein L10e
41	ABE	d.50.1	19	119	36	dsRNA-binding domain-like
42	ABE	d.55.1	19	119	36	Ribosomal protein L22
43	ABE	d.58.11	19	119	36	EF-G/eEF-2 domains III and V
44	ABE	d.66.1	19	119	36	Alpha-L RNA-binding motif
45	ABE	d.67.1	19	119	36	Threonyl-tRNA synthetase (ThrRS), second 'additional' domain
46	ABE	d.74.3	19	119	36	RBP11-like subunits of RNA polymerase
47	ABE	d.77.1	19	119	36	Ribosomal protein L5
48	ABE	e.24.1	19	119	36	Ribosomal protein L1
49	ABE	e.29.1	19	119	36	beta and beta-prime subunits of DNA dependent RNA-polymerase

A, B, E: respectively correspond to the number of species in Archaea, Bacteria and Eukarya that contain a given fold. The total number of species in the three superkingdoms are Archaea 19, Bacteria 119 and Eukarya 36.

Table 3.2 (B) Unique Folds exist exclusively within each region: E (5), AE (1), ABE (52)

Count	Region	Fold	A	B	E	SCOP definition
1	E	d.148	0	0	36	Ubiquitin-protein ligase E3a, Hect catalytic domain (E6ap)
2	E	d.189	0	0	36	PX domain
3	E	d.20	0	0	36	UBC-like

Table 3.2 (B) continued

Count	Region	Fold	A	B	E	SCOP definition
4	E	e.40	0	0	36	Cullin homology domain
5	E	g.45	0	0	36	Pyk2-associated protein beta ARF-GAP domain
1	AE	g.59	19	0	36	Zinc-binding domain of translation initiation factor 2 beta
1	ABE	a.156	19	119	36	S13-like H2TH domain
2	ABE	a.16	19	119	36	S15/NS1 RNA-binding domain
3	ABE	a.2	19	119	36	Long alpha-hairpin
4	ABE	a.27	19	119	36	Anticodon-binding domain of a subclass of class I aminoacyl-tRNA synthetases
5	ABE	a.4	19	119	36	DNA/RNA-binding 3-helical bundle
6	ABE	a.60	19	119	36	SAM domain-like
7	ABE	a.75	19	119	36	Ribosomal protein S7
8	ABE	b.34	19	119	36	SH3-like barrel
9	ABE	b.39	19	119	36	Ribosomal protein L14
10	ABE	b.40	19	119	36	OB-fold
11	ABE	b.43	19	119	36	Reductase/isomerase/elongation factor common domain
12	ABE	b.44	19	119	36	EF-Tu/eEF-1alpha/eIF2-gamma C-terminal domain
13	ABE	b.51	19	119	36	ValRS/IleRS/LeuRS editing domain
14	ABE	c.1	19	119	36	TIM beta/alpha-barrel
15	ABE	c.108	19	119	36	HAD-like
16	ABE	c.12	19	119	36	Ribosomal proteins L15p and L18e
17	ABE	c.120	19	119	36	PIN domain-like
18	ABE	c.2	19	119	36	NAD(P)-binding Rossmann-fold domains
19	ABE	c.20	19	119	36	Initiation factor IF2/eIF5b, domain 3
20	ABE	c.23	19	119	36	Flavodoxin-like
21	ABE	c.26	19	119	36	Adenine nucleotide alpha hydrolase-like
22	ABE	c.3	19	119	36	FAD/NAD(P)-binding domain
23	ABE	c.37	19	119	36	P-loop containing nucleoside triphosphate hydrolases
24	ABE	c.47	19	119	36	Thioredoxin fold
25	ABE	c.51	19	119	36	Anticodon-binding domain-like
26	ABE	c.55	19	119	36	Ribonuclease H-like motif
27	ABE	c.56	19	119	36	Phosphorylase/hydrolase-like
28	ABE	c.66	19	119	36	S-adenosyl-L-methionine-dependent methyltransferases
29	ABE	d.104	19	119	36	Class II aaRS and biotin synthetases
30	ABE	d.122	19	119	36	ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase
31	ABE	d.127	19	119	36	Creatinase/aminopeptidase
32	ABE	d.131	19	119	36	DNA clamp
33	ABE	d.14	19	119	36	Ribosomal protein S5 domain 2-like
34	ABE	d.140	19	119	36	Ribosomal protein S8
35	ABE	d.141	19	119	36	Ribosomal protein L6
36	ABE	d.142	19	119	36	ATP-grasp
37	ABE	d.15	19	119	36	beta-Grasp (ubiquitin-like)

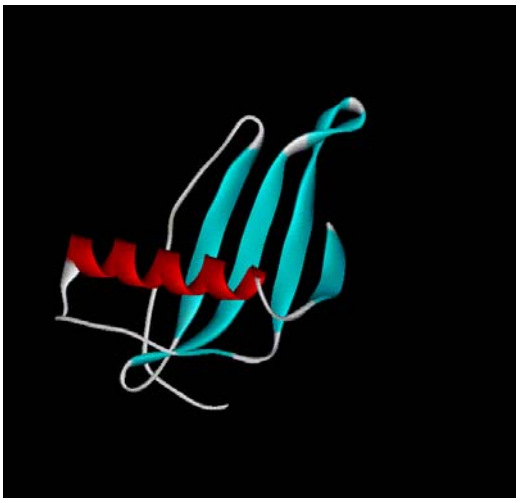
Table 3.2 (B) continued

Count	Region	Fold	A	B	E	SCOP definition
38	ABE	d.157	19	119	36	Metallo-hydrolase/oxidoreductase
39	ABE	d.26	19	119	36	FKBP-like
40	ABE	d.41	19	119	36	alpha/beta-Hammerhead
41	ABE	d.50	19	119	36	dsRBD-like
42	ABE	d.55	19	119	36	Ribosomal protein L22
43	ABE	d.58	19	119	36	Ferredoxin-like
44	ABE	d.66	19	119	36	Alpha-L RNA-binding motif
45	ABE	d.67	19	119	36	RRF/tRNA synthetase additional domain-like
46	ABE	d.68	19	119	36	IF3-like
47	ABE	d.74	19	119	36	DCoH-like
48	ABE	d.77	19	119	36	Ribosomal protein L5
49	ABE	d.79	19	119	36	Bacillus chorismate mutase-like
50	ABE	e.24	19	119	36	Ribosomal protein L1
51	ABE	e.29	19	119	36	beta and beta-prime subunits of DNA dependent RNA-polymerase
52	ABE	g.41	19	119	36	Rubredoxin-like

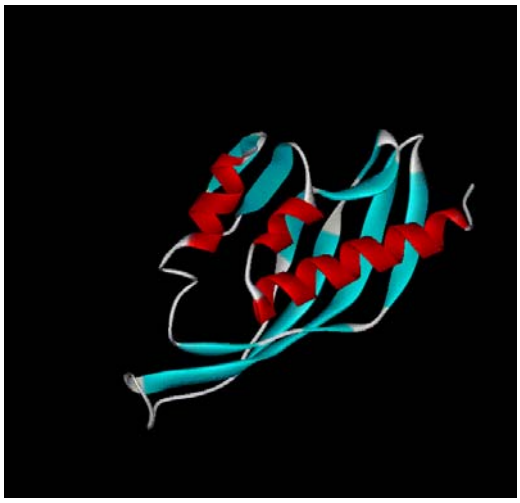
A single Domain Superfamily was identified that occurs in all 19 Archaea but nowhere else. The unique superfamily (SCOP superfamily d.17.6) is found in an enzyme involved in the synthesis of archaeosine, a modified base (7-formamidino-7-deazaguanosine) found exclusively in the Archaea [33]. The SCOP entry has been assigned to a fold level that includes six superfamilies, including cystatin-like folds. Direct visual comparison of the six types of structure shows that the archaeal domain in question is clearly distinctive from the other superfamilies in this fold group (Figure 3.11).

Figure 3.11 Structural comparison of the six Superfamilies in the d.17 Fold

d.17.1-1MOL



d.17.4-1IDP



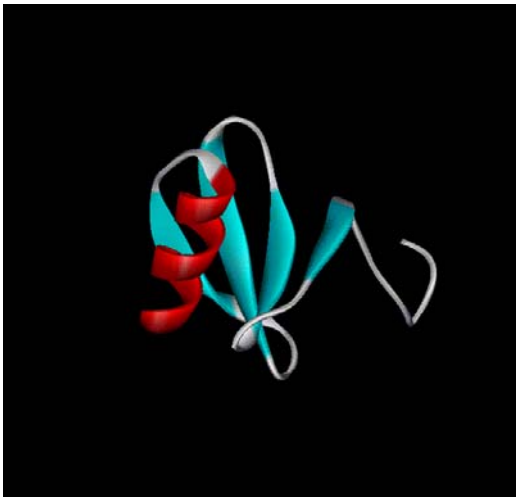
d.17.2-1OAC



d.17.5-1UDI



d.17.3-1EEJ



d.17.6-1IQ8



Disulfide Bonds

A typical three-dimensional structure of a protein domain is stabilized by large numbers of intramolecular interactions, such as hydrogen bonds and hydrophobic interactions. Small protein domains, with fewer numbers of these non-covalent interactions, are unstable and unlikely to exist. Disulfide bonds, the covalent bonds formed between two cysteine residues, in addition to other interactions, can stabilize small domains. Their existence likely contributed to the pool of protein domains. However, the disulfide bond itself is volatile under reducing conditions. Therefore, not until the earth's oxygen content were high enough to support the formation of disulfide bonds, the number of protein domains stabilized by disulfide bonds was not likely to be significant. Since the oxygen content of the earth's atmosphere has gradually increased during evolution, the emergence of disulfide bond-dependent domains may be correlated to the evolution of life. It has been estimated that Archaea, Bacteria and Eukaryota diverged about 1.8-2.2 billion years (byr) ago, and from 2byr to 1.5byr, the amount of oxygen in earth's atmosphere increased from insignificant to today's content. The process of emergence of disulfide bond-dependent domains can be reflected by their presents in each superkingdom (Figure 3.12).

The information about disulfide bonds is retrieved from PDB files of related protein domains. Although the domains having three-dimensional structures are only representatives of all of the protein domains having similar fold existing in all organisms, they still can be used as the first order estimation to investigate the content of disulfide bonds in general.

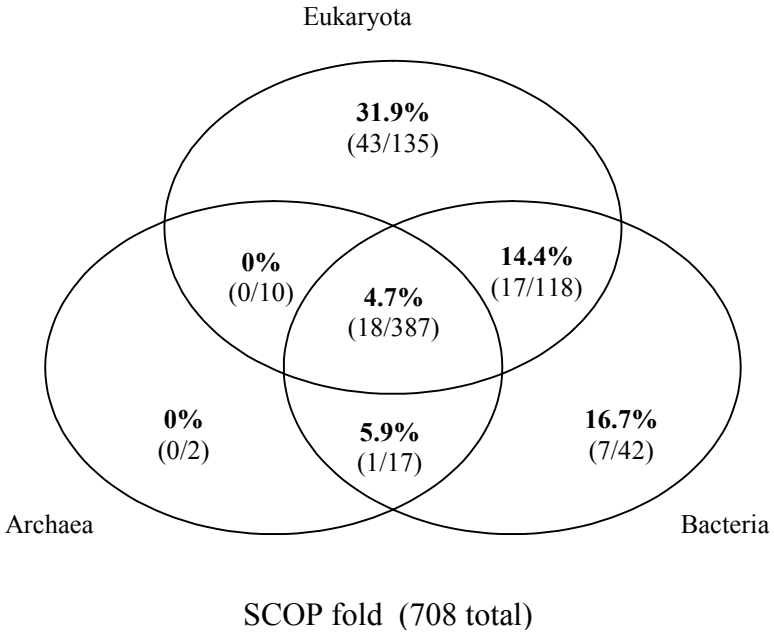


Figure 3.12 Numbers and percentages of disulfide bond-dependent domains in each kingdom. The criteria for disulfide bond-dependent is over 70% of the PDB representatives contain disulfide bond.

The result shown in Figure 3.12 largely confirms what we anticipated. Only 4.7% of the folds common to all superkingdoms contain disulfide bonds; most of the folds developed before the divergence of three kingdoms are not oxygen-dependent, and are prior to the accumulation of oxygen. On the contrary, 31.9% of the domains unique to Eukaryota are disulfide bond-dependent domains, which are developed when the required oxidizing environment was satisfied. Domains common to Bacteria and Eukaryotes have a smaller percentage (14.4%) of disulfide bond-containing folds, corresponding to their earlier time of development than Eukaryota domains. Correspondingly, Bacteria do not

have as many disulfide bond-dependent folds as Eukaryotes, probably because most of these folds function as intracellular molecules, which are significantly lacking in Bacteria. The percentage of disulfide bonds in AE (0%) and in AB (5.9%), which are close to that in ABE (4.7%), suggests they are likely to be of same origin as ABE, the common ancestor of all organisms.

3.4.2 Discussion

Superkingdom Fold Census

Past reports have represented the distribution of folds, shared and unique, in Venn diagrams similar to that depicted in Figure 3.10, although in those reports, the count was based on folds (as opposed to fold superfamilies) [13, 15] and were necessarily limited to fewer completed genomes. Nonetheless, the number of domains in the different sectors have continued to follow a similar pattern as more genomes have been reported and more folds identified. Most reports in the past have limited the census to any occurrence among a superkingdom's genomes. A more interesting number may be a count of those folds that are both unique and ubiquitous to a superkingdom or set of superkingdoms.

In this regard, the existence of about 50 folds that are common to all 174 genomes is a more important consideration than the overall abundance when it comes to the matter of common ancestors. It represents a core set of structures from which the most essential gene products are constructed.

Although these counts are not informative with regard to the detailed phylogenies, they do provide an interesting insight into related matters, especially when compared with those Superfamilies that occur uniquely and ubiquitously in any one of three superkingdoms (Figure 3.10 A). Although the 49 ubiquitous Superfamilies are mostly found in proteins involved in translation, including eight found in ribosomal proteins and six in aminoacyl-tRNA synthetases, several are clearly associated with major metabolic pathways. A full list of the 49 Superfamilies is included in Table 3.2.

That no unique and ubiquitous Superfamilies were found for Bacteria must be a reflection of their great diversity. The followup study with smaller numbers of bacterial and eukaryotic genomes showed that this finding was not merely attributable to the greater number of bacterial genomes available (Figure 3.10 B).

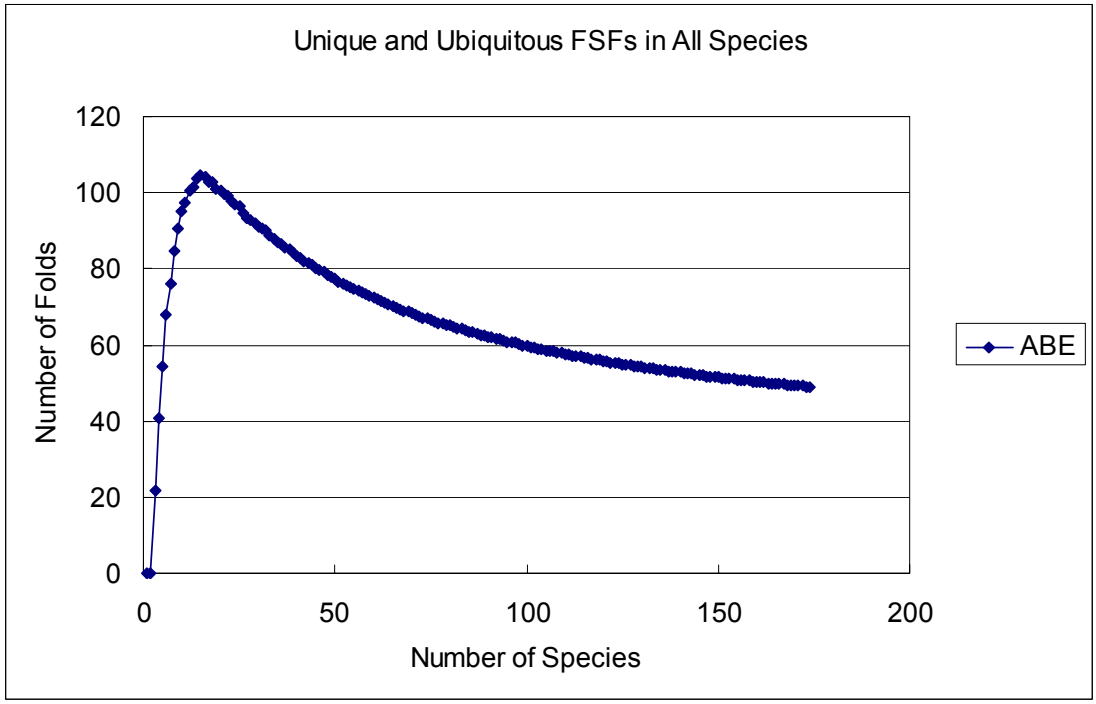
The single fold found in all 19 Archaea but nowhere else is noteworthy because its discovery underscores the advantage of using Superfamilies for phylogenetic studies. With only 19 Archaeal genomes completed, it would appear that a single (superfamily) fold, found in an enzyme responsible for modifying tRNAs, could be definitively diagnostic for membership in the Archaea (we are not suggesting such a narrow criterion). It will be of great interest to find if this fold is ever found to be absent in an archaeon or present in a bacterium or eukaryote.

The question arises as to whether enough data are in hand to make final judgments about fold occurrence and the overall phylogeny. More folds are likely to be discovered, and certainly more genomes will be sequenced. Interestingly, in a prior study made when there were only eight completed genomes, Gerstein [11] found that only 30 folds were

found in all of them. Now that we have identified about 50 such ubiquitous folds is not solely because new folds have been discovered. Inspection of Gerstein's data [11] reveal that many of the folds that have been added to the inventory were listed in that earlier report among those that occurred in 7 of the 8 genomes. What has improved in the interval is the sensitivity and certainty for correlating folds with ORFs in a genome. In this regard, currently >60% of the ORFs in the wholly sequenced genomes are being correlated with Superfamilies [25], whereas in early efforts such identifications amounted to <25% [11].

More genomes and better identifications not withstanding, it should be possible to estimate the ultimate number of Superfamilies on the basis of current data. To this end, the following exercise was conducted. The 174 genomes were sampled without replacement 1,000 times and the Superfamilies among them tallied to see how many were found in all chosen taxa. A smooth decay curve resulted (Figure 3.13 *Upper*), the curve-fitting for which was best matched by the general exponential expression.

Where x is the number of taxa used and y the number of ubiquitous Superfamilies. In the limit, when x approaches infinity, $y = 45.4$. By this reasoning, we feel that the number of ubiquitous Superfamilies is not likely to change radically in the future.



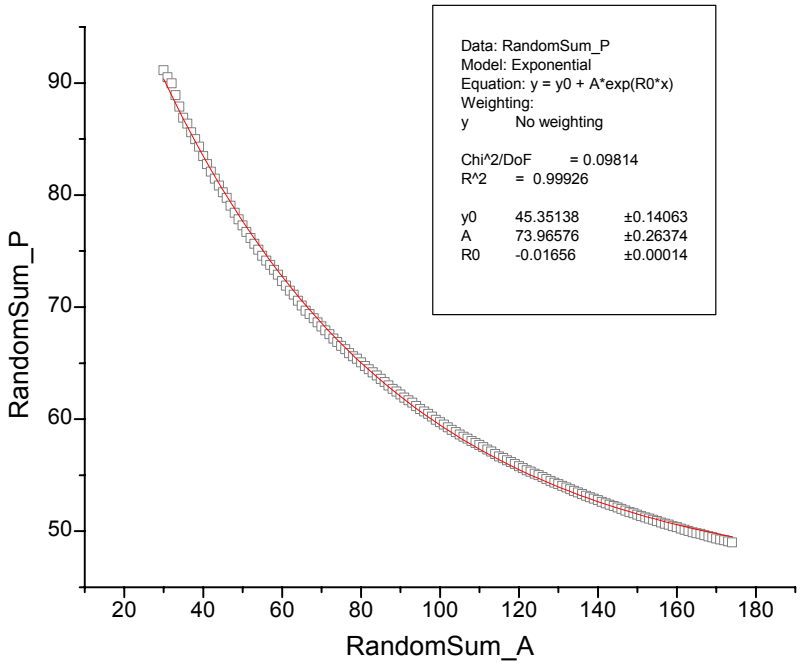


Figure 3.13 Simulation of the decay curve of the number of unique and ubiquitous protein domain superfamilies when all 174 genomes are randomly sampled without repeat 1000 times.

The Tree of Life

In recent years, the question has been raised as to whether there is a genuine tree of life with a common ancestral organism at its root [34-37]. Instead, challengers propose that the three superkingdoms arose independently from a community of primitive and undifferentiated cells that were freely exchanging their genetic components. What the three lineages shared, and what gives rise to the trichotomy provided by ribosomal RNA comparisons, was a common translation machinery but not much else. Woese [37] refers to the transition point at which vertically transmitted genes began to weigh more heavily than horizontally transferred ones as the Darwinian Threshold. The transition was marked, he feels, by a level of "componentry" that for organizational reasons could no longer survive by reckless and promiscuous exchange. He posits that the threshold came well after the three realms of life had emerged.

We feel that the census of folds sheds light on the matter. The discovery that there are 49 superfamily folds common to all 174 genomes seems to us to argue that there was a last common ancestor for all three superkingdoms, and that it had a very sophisticated genetic inventory of structural equipment, as represented by the 49 different domains being found in a wide range of gene products over and beyond those having to do with translation.

3.5 Conclusion

In summary, a simple tallying of Superfamilies for 174 fully sequenced genomes has been used for constructing phylogenies that are in good accord with those based on ribosomal RNA sequences and other genetic information. The detailed distribution of folds among the Archaea, Bacteria, and Eukaryote is further revealing about the evolutionary history of life on Earth and may ultimately prove useful for detailing the routes of their diversification. Recent studies showing how structural differences in folds are only compatible with a scenario of divergence [38] underscore the promise of the approach.

Chapter 3 contains the material as it appears in Yang, S., R.F. Doolittle, and P.E. Bourne, Phylogeny determined by protein domain content. Proc Natl Acad Sci U S A, 2005. 102(2): p. 373-8. The dissertation author was the primary investigator and author of this paper.

3.6 References

1. House, C.H. and S.T. Fitz-Gibbon, *Using homolog groups to create a whole-genomic tree of free-living organisms: an update.* J Mol Evol, 2002. **54**(4): p. 539-47.

2. Wolf, Y.I., et al., *Genome trees and the tree of life.* Trends Genet, 2002. **18**(9): p. 472-9.

3. Tekaia, F., A. Lazcano, and B. Dujon, *The genomic tree as revealed from whole proteome comparisons.* Genome Res, 1999. **9**(6): p. 550-7.

4. Bansal, A.K. and T.E. Meyer, *Evolutionary analysis by whole-genome comparisons.* J Bacteriol, 2002. **184**(8): p. 2260-72.

5. Brown, J.R., et al., *Universal trees based on large combined protein sequence data sets.* Nat Genet, 2001. **28**(3): p. 281-5.

6. Baldauf, S.L., et al., *A kingdom-level phylogeny of eukaryotes based on combined protein data.* Science, 2000. **290**(5493): p. 972-7.

7. Snel, B., P. Bork, and M.A. Huynen, *Genome phylogeny based on gene content.* Nat Genet, 1999. **21**(1): p. 108-10.

8. Dandekar, T., et al., *Conservation of gene order: a fingerprint of proteins that physically interact.* Trends Biochem Sci, 1998. **23**(9): p. 324-8.

9. Korbel, J.O., et al., *SHOT: a web server for the construction of genome phylogenies.* Trends Genet, 2002. **18**(3): p. 158-62.

10. Wolf, Y.I., et al., *Genome trees constructed using five different approaches suggest new major bacterial clades.* BMC Evol Biol, 2001. **1**: p. 8.

11. Gerstein, M., *Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census.* Proteins, 1998. **33**(4): p. 518-34.

12. Gerstein, M. and H. Hegyi, *Comparing genomes in terms of protein structure: surveys of a finite parts list.* FEMS Microbiol Rev, 1998. **22**(4): p. 277-304.

13. Wolf, Y.I., et al., *Distribution of protein folds in the three superkingdoms of life.* Genome Res, 1999. **9**(1): p. 17-26.

14. Lin, J. and M. Gerstein, *Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels.* Genome Res, 2000. **10**(6): p. 808-18.

15. Caetano-Anolles, G. and D. Caetano-Anolles, *An evolutionarily structured universe of protein architecture.* Genome Res, 2003. **13**(7): p. 1563-71.

16. Bapteste, E., et al., *Phylogenetic reconstruction and lateral gene transfer.* Trends Microbiol, 2004. **12**(9): p. 406-11.

17. Clarke, G.D., et al., *Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores.* J Bacteriol, 2002. **184**(8): p. 2072-80.

18. Roelofs, J. and P.J. Van Haastert, *Genes lost during evolution.* Nature, 2001. **411**(6841): p. 1013-4.

19. Rokas, A., et al., *Genome-scale approaches to resolving incongruence in molecular phylogenies.* Nature, 2003. **425**(6960): p. 798-804.

20. Kunin, V. and C.A. Ouzounis, *The balance of driving forces during genome evolution in prokaryotes.* Genome Res, 2003. **13**(7): p. 1589-94.

21. Krylov, D.M., et al., *Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution.* Genome Res, 2003. **13**(10): p. 2229-35.

22. Gogarten, J.P., W.F. Doolittle, and J.G. Lawrence, *Prokaryotic evolution in light of gene transfer.* Mol Biol Evol, 2002. **19**(12): p. 2226-38.

23. Snel, B., P. Bork, and M.A. Huynen, *Genomes in flux: the evolution of archaeal and proteobacterial gene content.* Genome Res, 2002. **12**(1): p. 17-25.

24. Gough, J., et al., *Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.* J Mol Biol, 2001. **313**(4): p. 903-19.

25. Gough, J. and C. Chothia, *SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments.* Nucleic Acids Res, 2002. **30**(1): p. 268-72.

26. Felsenstein, J., Cladistics, 1989. **5**: p. 164-166.

27. Sokal, R.R. and C.D. Michener, Univ. Kans. Sci. Bull., 1958. **28**: p. 1409-1438.

28. Saitou, N. and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Mol Biol Evol, 1987. **4**(4): p. 406-25.

29. Waters, E., et al., *The genome of Nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism.* Proc Natl Acad Sci U S A, 2003. **100**(22): p. 12984-8.

30. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J Mol Biol, 1995. **247**(4): p. 536-40.

31. Lo Conte, L., et al., *SCOP database in 2002: refinements accommodate structural genomics.* Nucleic Acids Res, 2002. **30**(1): p. 264-7.

32. Doolittle, R.F., *The multiplicity of domains in proteins.* Annu Rev Biochem, 1995. **64**: p. 287-314.

33. Ishitani, R., et al., *Crystal structure of archaeosine tRNA-guanine transglycosylase.* J Mol Biol, 2002. **318**(3): p. 665-77.

34. Kandler, O., *CELL-WALL BIOCHEMISTRY AND 3-DOMAIN CONCEPT OF LIFE.* Syst. Appl. Microbiol., 1994. **16**: p. 501-509.

35. Koga, Y., et al., *Did archaeal and bacterial cells arise independently from noncellular precursors? A hypothesis stating that the advent of membrane phospholipid with enantiomeric glycerophosphate backbones caused the separation of the two lines of descent.* J Mol Evol, 1998. **46**(1): p. 54-63.

36. Woese, C., *The universal ancestor.* Proc Natl Acad Sci U S A, 1998. **95**(12): p. 6854-9.

37. Woese, C.R., *On the evolution of cells.* Proc Natl Acad Sci U S A, 2002. **99**(13): p. 8742-7.

38. Deeds, E.J., B. Shakhnovich, and E.I. Shakhnovich, *Proteomic traces of speciation.* J Mol Biol, 2004. **336**(3): p. 695-706.

Chapter 4 Genome Alignment Using Protein Domain

4.1 Introduction

In the era of complete genomes, the sequences of whole genomes or chromosomes can be studied and compared, enabling the study of the evolution of genome structure and architecture [1, 2]. Previous work has shown that various evolutionary events, such as inversion, translocation, duplication, as well as insertion and deletion, frequently occurred during evolution and altered the genomic structure of an organism [3]. It is found that many features of the evolution of chromosomes resemble that of proteins, although at different scales (Table 4.1).

Table 4.1 Comparison of evolution of protein sequence and chromosome architecture.

	Protein	Chromosome
Basic element	Amino acid	Genes (or protein domains)
Common Evolutionary events	Insertion, deletion, duplication	Horizontal gene transfer, deletion, duplication
Unique Evolutionary events	Substitution	Inversion, translocation
Constraint	Protein structure and function	Little constraint on genome architecture

Similar to protein sequence alignment, which is used to determine the evolutionary relationship of proteins based on their primary amino acid sequence, we proposed an approach that uses genome alignment to compare the relationship of

genomes using their gene or protein domain sequence on the chromosomes. As the fundamental evolutionary units, protein domains are considered the basic unit of chromosome, and the comparison of chromosomes are based on the identity and similarity of protein domains.

Similar approaches based on gene positions in chromosomes have been tested before [4-6]. The difference and advantages of our approach based on protein domain sequences will be discussed later.

4.2 Method

4.2.1 Data Management

The SUPERFAMILY database release 2004-10-05 provides the domain assignment for each protein, whose position in chromosomes is retrieved from genome mapping data. Specifically, complete genome sequence of over 300 organisms were obtained from several genome sequence centers, such as the National Center for Biotechnology Information (NCBI) [http://www.ncbi.nlm.nih.gov/Genomes/], the TIGR Comprehensive Microbial Resource [www.tigr.org/], and the ENSEMBL Genome Center [www.ensembl.org/], etc. The position of each protein and domain on chromosomes are derived accordingly.

4.2.2 Domain Dot Plot

Protein domain representation of chromosomes is constructed by sequentially assembling the predicted protein domains according to their positions on the chromosomes. Each domain is represented by one dot; unassigned parts within proteins and other regions in the chromosomes, such as repetitive elements and regulatory elements, are not listed. Similar to pairwise sequence comparison, protein domain sequences of chromosomes from two organisms are compared using a dot plot. In the dot plot, the horizontal and vertical axes represent the chromosomes of the two species being compared, and each pixel on the axes is a protein domain at the superfamily level. A dot

on the map indicates that the two species have the same domain at the corresponding position on the chromosome. In sequence comparison, because the types of elements are limited, four nucleotides for DNA or 20 amino acids for protein, a comparison window of size 4-11 is used to improve signal-to-noise ratio. This is not necessary when using protein domains.

4.2.3 Dynamic Programming

Similar to sequence comparison, dynamic programming has been used to find the correlated domain pairs in two chromosomes. As shown in a later section, inversions of part of the chromosome are common evolutionary events, which will cause extensively fragmented and inverse alignments. Therefore, in genome alignment, instead of global alignment, local alignment is used to pick up the scattered alignment segments. In addition, inverse chromosome alignments are performed to identify all the chromosome inversions.

4.3 Results

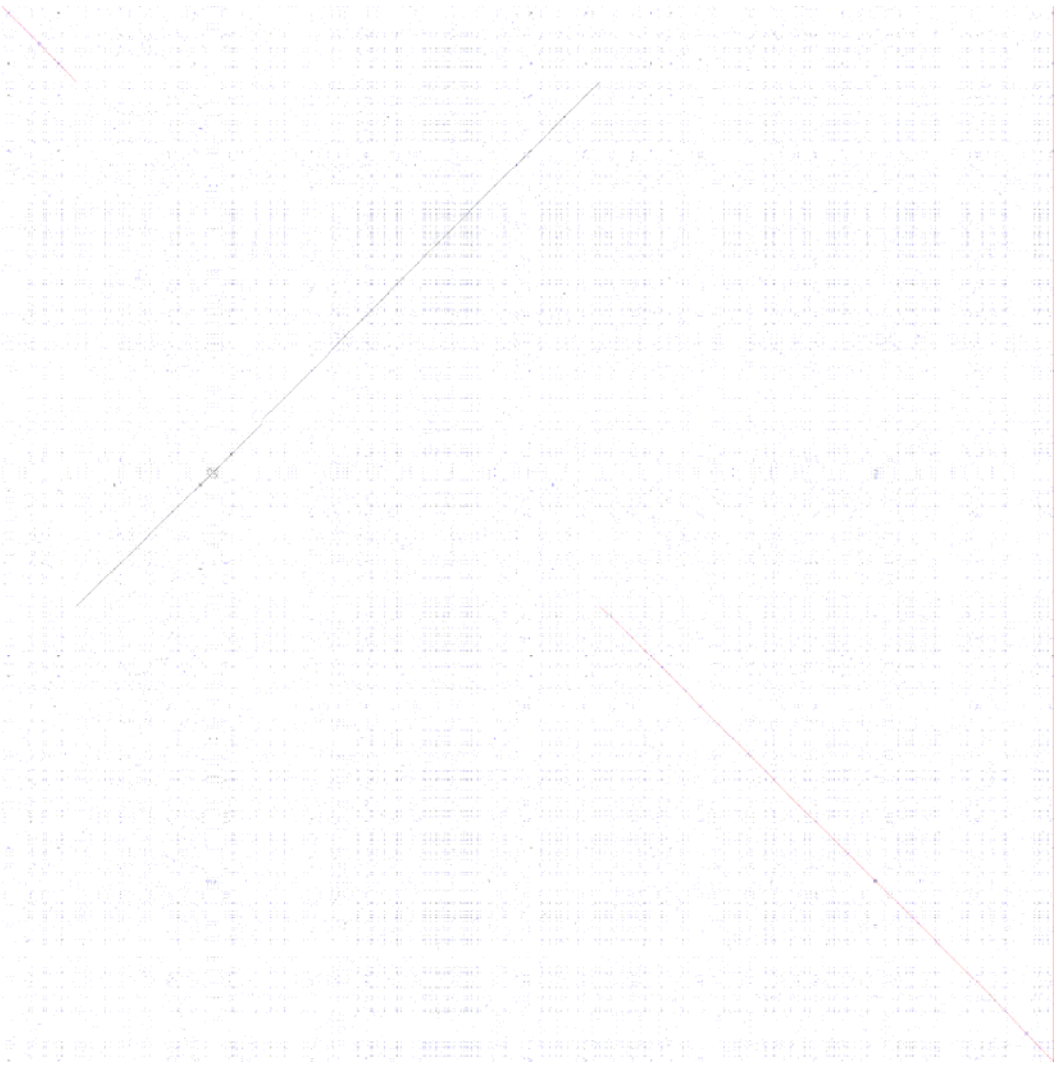
4.3.1 Genome Evolution Viewed by Domain Dot Plots

Figure 4.1 shows the dot plot for two closely related organisms, two strains from the same species, *Salmonella enterica*, *CT18* and *Ty2*. *Salmonella enterica CT18* has 4395 proteins in a single circular chromosome, and SUPERFAMILY predicted 3467 protein domains; *Salmonella enterica Ty2* is about the same size. The domain alignments (red dots) along the diagonal of the plot indicate that the majority of the chromosomes from the two strains are aligned, except a major inversion occurs in the first half of the chromosome. This chromosome inversion was previously discovered by whole genome comparison based on DNA sequences, which also confirms the validity of our genome alignment method based on protein domains to identify the major genomic rearrangement events [7].

Besides inversion, translocations of chromosomes occur frequently during evolution, as can be seen in the comparison of two *Chlamydophila pneumoniae* strains (Figure 4.2). Translocations usually move the domain alignment away from the diagonal of the plot, but do not change the orientation of the alignment; whereas insertions have the opposite effect on the dot plot.

Besides local duplication, we haven't observed whole genome duplication or duplication of large fragments of the chromosome, probably because the genome duplicates diverge very fast after the duplication events and there are no two organisms in our dataset that are evolutionarily close enough to illustrate the duplication events.

A



B

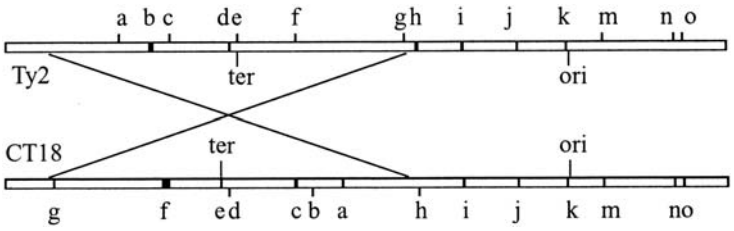


Figure 4.1 (A) Domain dot plot of two *Salmonella enterica* strains, *CT18* (vertical) and *Ty2* (Horizontal). Blue dots represent domain matches, red dots represent the calculated domain local alignment and green dots represent extrapolated local alignment. (B) The chromosome inversion determined by comparative genome analysis [7].

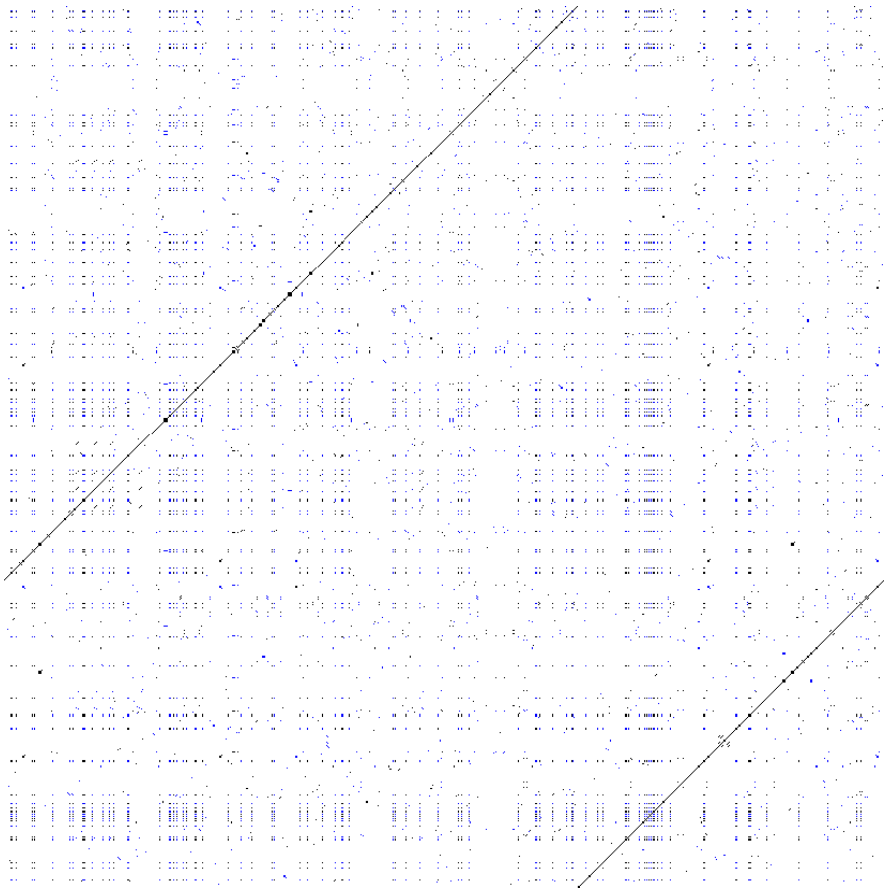
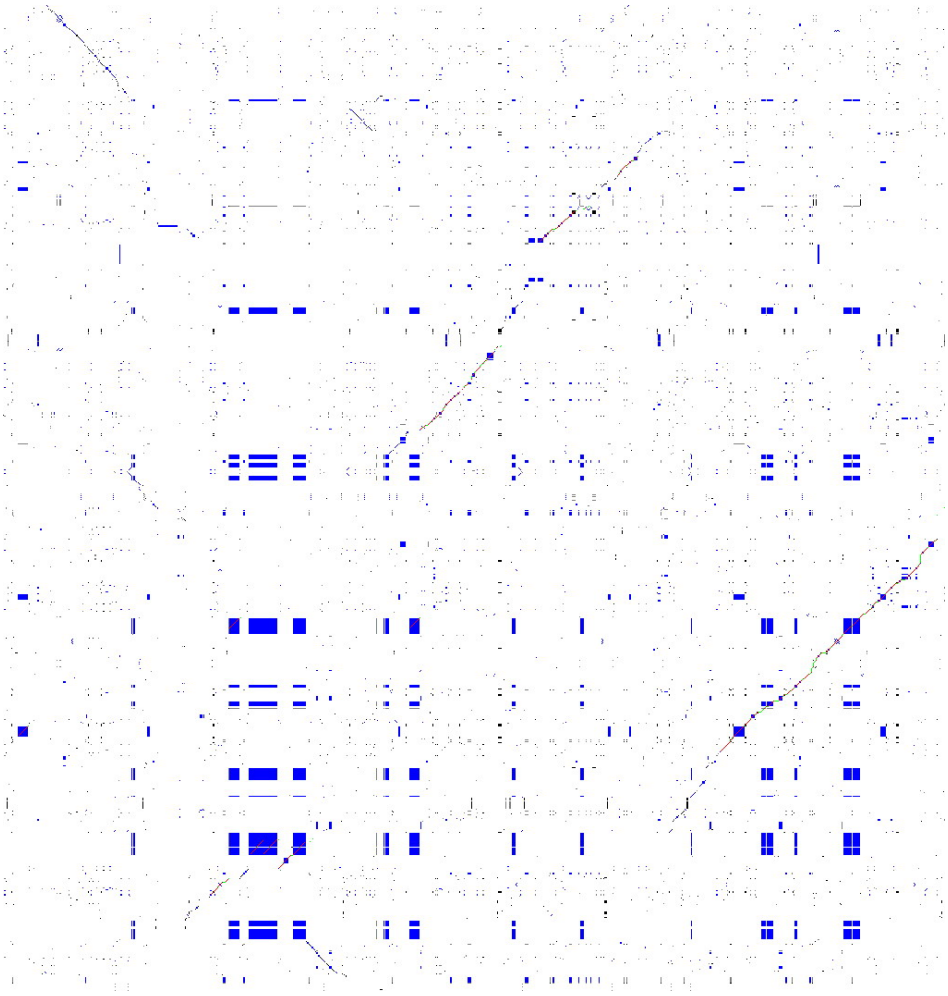


Figure 4.2 Domain dot plot for Chlamydophila pneumoniae AR39 and CWL029 represents a major chromosome translocation.

Other evolutionary events such as insertion, deletion and horizontal transfer are represented by gaps in one of the chromosomes being compared. As one can imagine, when the two species become more distant, the inversion/translocation as well as insertion/deletion/transfer will be less prevalent and the genome alignment will be less resolved, as can be seen in the comparison of the X chromosomes of human and mouse (Figure 4.3).

A



B

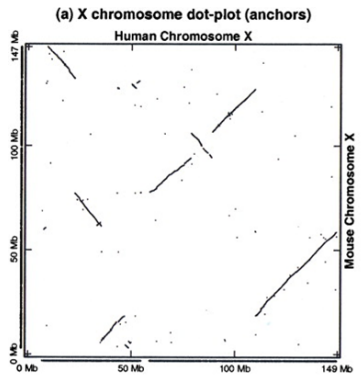
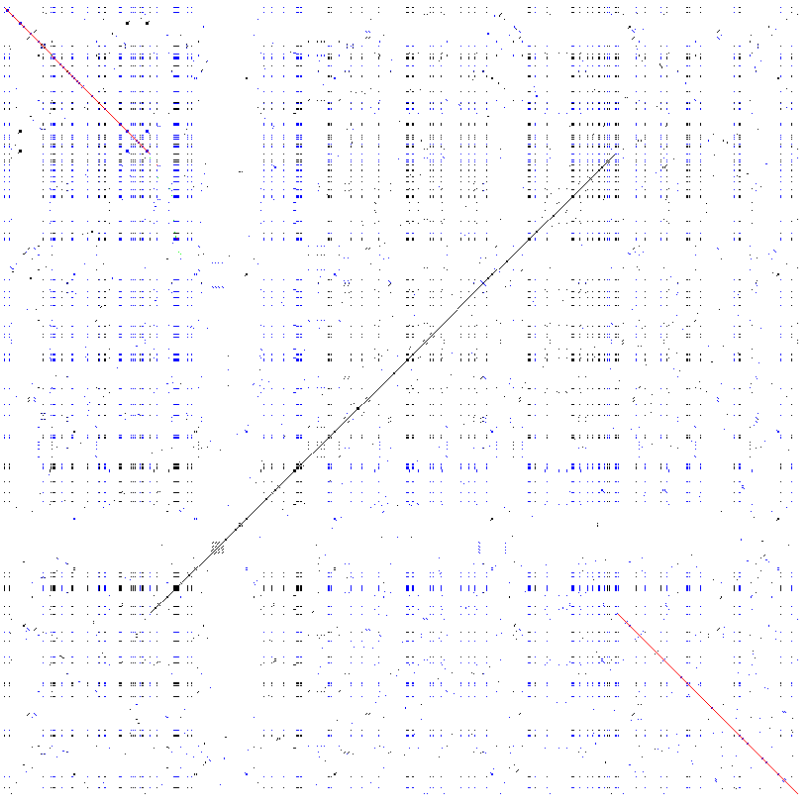


Figure 4.3 Comparison of the X chromosomes of human and mouse by (A) Protein domain comparison. (B) Genome comparison. Boxes in the plot represent proteins with tandem repeat domains, which exist extensively in higher eukaryotes [8].

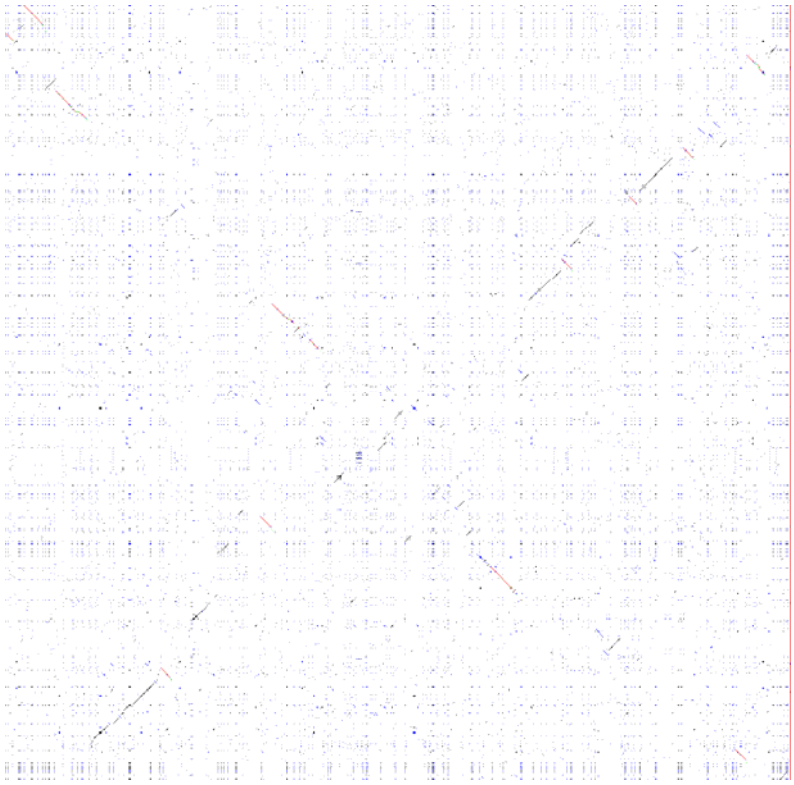
4.3.2 Symmetrical inversion of prokaryote genomes around the replication origin

Using our protein domain-based genome comparison approach, we found that many prokaryotic genomes experienced one or multiple steps of partial inversion during evolution and all the inversions were centralized approximately at the origin of replication (Figure 4.4). (In a circular chromosome symmetrical inversion around the origin and the terminus of replication is equivalent) The resulting genome alignment dot plot becomes X-shaped. Figure 4.5 schematically illustrates the formation of the X-shape alignment in dot plot after multiple steps of chromosome inversion around the replication origin.

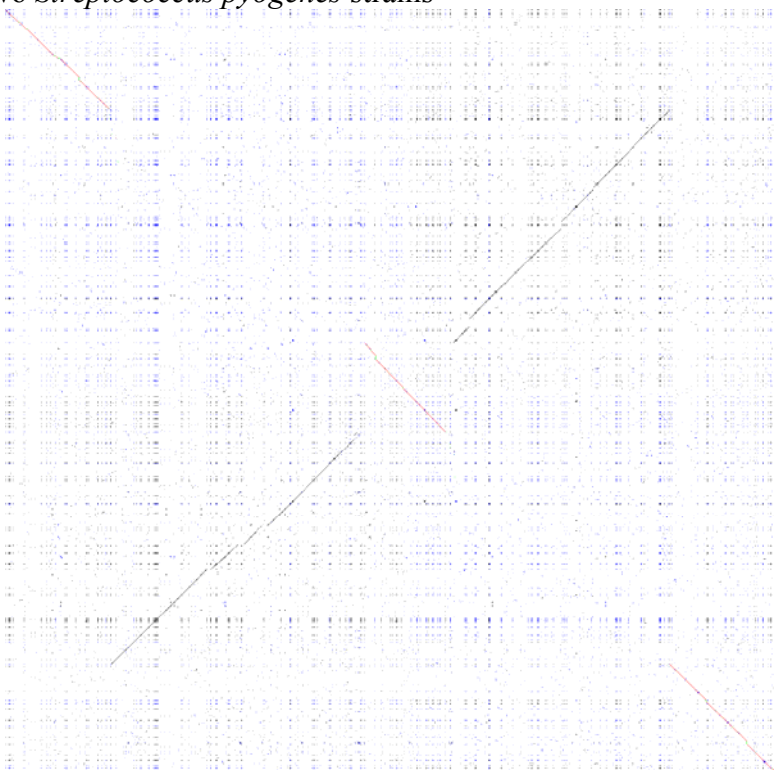


(A) Two *Tropheryma whipplei* strains

Figure 4.4 The genome alignment of three pairs of closely related species, each pair involves various genome inversion.



(B) Two *Streptococcus pyogenes* strains



(C) *Thermoplasma acidophilum* and *Thermoplasma volcanium*

Figure 4.4 Continued

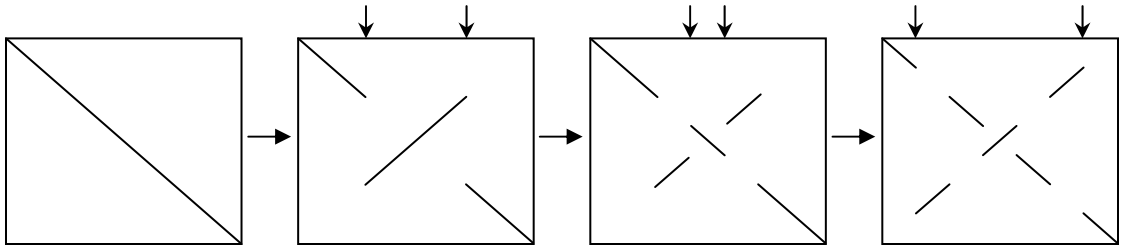


Figure 4.5 Multiple chromosome inversions arround the replication origin leads to the X-shape genome alignment observed in the dot plot. Replication origins of the circular chromosome are located at the start position of the plot. Arrows at the top of each plot mark the location of inversion.

The symmetrical inversion around the replication origin in prokaryotic genomes has been previously discovered using gene position plots [4-6]. This pattern has been thought to be the result of high recombination frequency at the open replication forks. Without prior knowledge of their previous result, we rediscovered the same pattern using our domain dot plot method.

4.4 Discussion

4.4.1 Domains as Evolutionary Units

In chapter 3, the presence and absence of protein domains in complete genomes are used to reconstruct the phylogenetic relationship of species from all major taxa of the three superkingdoms. In this chapter, protein domains are considered as the basic unit of the chromosomes and genomes, and the domain sequences are used to compare the whole chromosome or genomes of closely related organisms. Not only the abundance and type of protein domains, but also the positions of domains along the chromosomes reflect the evolutionary processes that genomes are subjected to.

Unlike DNAs that have four types of nucleic acids, and proteins that have twenty amino acids, protein structure space currently contains over 2000 protein domain families, a number large enough that the probability of mismatch in chromosome alignment is very low. Therefore, whole genome comparison, which is usually based on the nucleic acid sequence, can be improved both in accuracy and in efficiency if protein domains, the basic evolutionary units of genomes, are first identified.

On the other hand, compared with the total protein repository, the number of protein domains is limited. Genome alignment by genes must also require the correct identification of orthologs and paralogs of proteins, which, as discussed in chapter 3, is difficult. In addition, protein domain families do not distinguish ortholog and paralog, enabling the recognition of duplication events in each chromosome.

In summary, chromosome alignment by protein domains is able to discover major evolutionary events identified by gene position plots, with simple data management and calculation, low signal-to-noise ratio and clear visualization. This method can potentially provide more information about genome evolution and domain evolution.

4.4.2 Future Directions

Because of extensive chromosome inversion and translocation, genome alignment is best applied to closely related organisms, typically within the same Family in taxonomic group. The genome architecture of different organisms can be used to derive phylogenetic relationships among species, as well as the time when the evolutionary events, such as inversion and translocation, took place.

Evolution of domains, such as divergence of domains into two families, or emergence of new domains, might also be identified by comparing closely related genomes. Genomic positions of domains provide additional knowledge of the relationship between domains besides amino acid sequence; therefore, future efforts should be focusing on this topic.

Coevolution and coexpression are two areas to which this technique could be applied. For example, groups of genes that are coexpressed could show evidence of related colocolization on the chromosome/genome.

4.5 References

1. Miller, W., et al., *Comparative genomics.* Annu Rev Genomics Hum Genet, 2004. **5**: p. 15-56.

2. Bentley, S.D. and J. Parkhill, *Comparative genomic structure of prokaryotes.* Annu Rev Genet, 2004. **38**: p. 771-92.

3. Mira, A., L. Klasson, and S.G.E. Andersson, *Microbial genome evolution: sources of variability.* Current Opinion In Microbiology, 2002. **5**(5): p. 506-512.

4. Tillier, E.R.M. and R.A. Collins, *Genome rearrangement by replication-directed translocation.* Nature Genetics, 2000. **26**(2): p. 195-197.

5. Eisen, J.A., et al., *Evidence for symmetric chromosomal inversions around the replication origin in bacteria.* Genome Biol, 2000. **1**(6): p. RESEARCH0011.

6. Suyama, M. and P. Bork, *Evolution of prokaryotic gene order: genome rearrangements in closely related species.* Trends Genet, 2001. **17**(1): p. 10-3.

7. Deng, W., et al., *Comparative genomics of Salmonella enterica serovar Typhi strains Ty2 and CT18.* J Bacteriol, 2003. **185**(7): p. 2330-7.

8. Pevzner, P. and G. Tesler, *Genome rearrangements in mammalian evolution: lessons from human and mouse genomes.* Genome Res, 2003. **13**(1): p. 37-45.

Chapter 5 Evolution of Protein Domains

5.1 Introduction

5.1.1 The Evolutionary History of Protein Domains

In previous chapters, protein domain content has been proven to be suitable to decipher the phylogenetic relationships among species in the tree of life, underscoring the idea that protein domains are fundamental evolutionary units. Indeed, the distribution of various protein domains has been used to answer fundamental evolutionary questions, such as the origin of eukaryotes and the divergence of the three superkingdoms.

On the other hand, the evolution of each protein domain and the formation of the protein domain repertoire are also of great interest to scientists. Questions such as the evolutionary origins of protein domains, identification of protein domain loss, transfer, duplication and combination with other domains to form new proteins, and the influence of the evolution of domains on the evolution of proteins and functions as well as the organisms as a whole, are fundamental and challenging topics in evolutionary biology.

The evolution of protein domains consists of two different but related aspects: the changes to protein domains themselves and changes in domain organization within the complete genome. The former includes important events such as the innovation of new domains and the formation of new domain combinations. There are several hypotheses concerning the origin of protein domains. One hypothesis claims that most of domains come from duplication of previous domains. New domains form through either

divergence of one of the duplicates or from a novel combination of the duplicates. Other possibilities might include *de novo* genesis of domain from scratch. Recombination of existing domains is a major source of new proteins and new functions through the. Although being involved in a new combination will not change a domain immediately, it definitely changes the structural environment and the evolutionary constraint of the domain, thus will eventually affect its sequence, structure and function. Therefore, domain combinations are considered changes of domains themselves.

On the other hand, evolutionary events, such as duplication, loss, and transfer between species will change the genomic content of domains or domain combinations without changing their identities. These two aspects of domain evolution are correlated. The emergence of a new domain in a species not only depicts the origin of the domain but also indicates its existence in the genome. The duplication of a domain may induce the divergence through mutations, insertions and deletions, producing modified structures and functions that distinguish it from its ancestor.

A big challenge is understanding the evolutionary events that took place during the past 3.8 billion years involving over 1000 unique folds, and their numerous combinations. The sequencing of complete genomes across the major lineages of the tree of life provides an opportunity to uncover the entire evolutionary history of protein domains.

5.1.2 Previous Work

As mentioned in Chapter 1, in the last ten years, with the accumulation of complete genome and the improvement of homology detection algorithms, people started to investigate the distribution of protein domains in the three superkingdoms of life and other aspects of protein domain evolution [1-7].

Cyrus Chothia, Sarah Teichmann and their colleges in MRC are one of the most active groups in this area. One of the first observations involves the existence of multi-domain proteins in the three superkingdoms of life. Rough estimate indicates that two-thirds of prokaryote proteins have two or more domains, whereas the percentage of multi-domain proteins in Eukaryotes increases to about four-fifth [8]. In 2001, the investigation of domain combinations in 40 genomes showed a power-law distribution of the tendency of domains to form combinations [9]. The authors also observed conservation of domain orientations. In another paper, Christine Vogel, et al. described two-domain or three-domain combinations that frequently recur in different protein contexts, which were called “supra-domains” because of their larger sizes [10]. A simulation of the processes of domain duplication and combination suggests that domain combinations are stochastic processes followed by duplication to various extents [11]. Sarah Kummerfeld, et al. investigated the relative rates of gene fusion and fission events in multi-domain proteins using phylogenetic information [12]. They estimated that fusion is four time more common than fission. In another attempt, Julian Gough tried to evaluate the extent of convergent evolution of domain architectures, which in principle exist but were hard to prove [13]. He claimed that the chance of convergent evolution through domain

combination is rare, approximately 0.4-4%. A recent analysis suggested that the abundance of protein domains and domain combinations are correlated with the complexity of the organism, as characterized by the numbers of cell types each organism contains [14].

Christine Orengo and colleagues approach the topic of domain combination using the CATH [15, 16] protein classification scheme and the Gene3D [17] genomic domain assignment database. They found a correlation of the domain abundance and genome size [18]. The abundance of size-independent domains is not correlated with increase in genome size, whereas the occurrence of size-dependent domains does correlate with genome size and the correlation can be further grouped into linear and non-linear. The three groups of domains, the size-independent, linearly distributed and non-linearly distributed, are roughly functionally involved in protein translation and biosynthesis, metabolism, and gene regulation, respectively. Arne Elfsson and his group also make contribution in domain recombination [19, 20] and duplication [21].

In addition, there are several theoretical attempts based on mathematical modeling and physical simulation that aim to derive the evolutionary history of proteins and hence protein structure space, especially relating to the origin of domains and their presence in the universal last common ancestor [22-26].

5.1.3 The Hypothesis to Be Tested

Previous approaches are commonly based on the distribution of protein domains in complete genomes across the three superkingdoms. The over one hundred (now over 300) organisms were roughly divided into three superkingdoms, Archaea, Bacteria and Eukaryotes, and only common features at the superkingdom level were compared and analyzed. The Venn diagram analysis of domain distribution in Chapters 3 is one example. In these studies, the phylogenetic relationships among species across all the major lineages of the tree of life have not been fully appreciated and the huge differences in genotype and phenotype among the species in each superkingdom have not been considered.

Here we propose an approach that takes full advantage of existing phylogenetic information for each species to derive the entire evolutionary history of each domain throughout the tree of life. In fact the three superkingdoms in the Venn diagram only represent the three branches of the phylogeny at the root of the tree (Figure 5.1). The distribution of domains in the three superkingdoms reflects evolutionary process during the divergence from the last universal common ancestor into the three superkingdoms. Our approach aims to generalize such an analysis to every node in the entire tree of life.

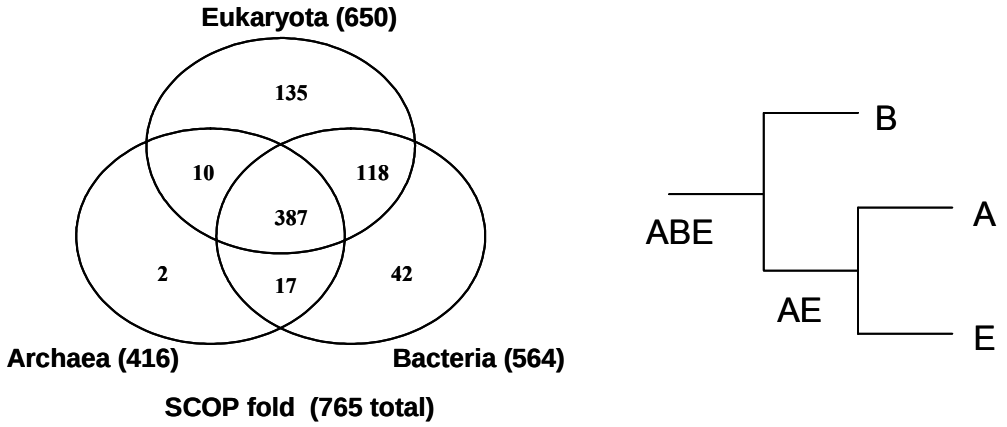


Figure 5.1 The Venn diagram of the distribution of folds in the three superkingdoms of life (left) and the corresponding phylogenetic tree at the root of the life (right).

Previously the standard procedure in studying the evolution of a single protein family typically starts with the identification of all sequence homologs to that protein family through a BLAST search against all known sequences in GenBank. Then a gene tree generated by a multiple sequence alignment is compared to the species tree. By using the phylogeny of the organisms that are found to contain this protein, the evolutionary origin of genes, gene duplication, loss, horizontal gene transfer and other evolutionary events can be identified. Our approach is similar to the above procedure when one protein or domain family is considered. Indeed, because of the pre-computed genomic domain assignment data, our approach can automatically investigate the evolutionary history of all domains in the protein domain universe concurrently.

5.2 Domain Tree and Domain Combination Tree

5.2.1 Methods

Data Source

Genomic domain assignments were retrieved from SUPERFAMILY release 2005-10-19, which includes 23 Archaea, 233 Bacteria and 59 Eukaryotes. In this release not only did the number of complete genomes increase but further the domain assignment has been calculated at the family level. We use the same e-value cutoff as previous chapters when handling the data.

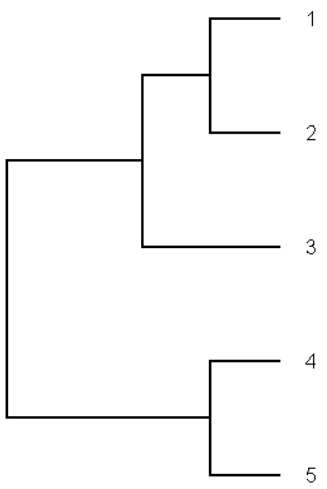
Standard Phylogeny and NCBI Taxonomy

We consider the NCBI taxonomy as the standard phylogeny of the tree of life. There are subtle distinctions between taxonomy and phylogeny, where biological taxonomy is the classification and naming of biological organisms and phylogeny refers to the evolutionary history of the organisms represented by a tree. However, taxonomy lists the unique characters of each taxon and is designed to reflect phylogeny [27]. The molecular phylogenetic methods based on sequence homology have lead to fundamental discoveries in phylogenetics as well as dramatic changes in taxonomy.

We retrieved the taxonomy plain text files from NCBI, and only those selected the species that have complete genomes in our Superfamily dataset. We subsequently

created tree files in a standard format, which can be interpreted and visualized by various tree-viewing packages (Figure 5.2).

A



B

(((1,2),3),(4,5));

Figure 5.2 (A) a schematic phylogenetic tree represented by (B) a text file in the standard tree format. The tree is generated by the TreeView software.

An intrinsic problem in taxonomy-based phylogeny is that the tree is not a binary tree in which every node only contains two leaves. The hierarchical classification usually consists of six levels: kingdom, class, order, family, genus and species. Each level / node contains multiple sublevels / leaves; hence, the taxonomy-based tree is a not a binary tree, which a phylogenetic tree should be in theory. The discrepancies between the taxonomy-

based phylogeny and the actual phylogeny as well as its impact on our analysis will be discussed later.

It's worth noting that beside the NCBI taxonomy-based tree other phylogenetic trees of life can be used in this analysis, including the protein domain content-based tree (Chapter 3). The NCBI taxonomy-based tree is chosen because it is considered the gold standard of phylogeny. The NCBI taxonomy is based on extensive genetic and morphological evidence, and is built by standard and molecular phylogenetic methods. One basic assumption of this approach is that we have a correct tree of life. Even if our tree is not perfect we can learn a great deal. This understanding will only improve as the tree is further refined.

Domain Tree

The phylogenetic tree based on taxonomy illustrates the evolutionary history of the organisms under study. The leaves of the tree represent the over 300 currently existing species; whereas each internal node represents the most recent common ancestor of its descendants. Internal nodes are generally referred as Hypothetical Taxonomic Units (HTU) since they can not be observed today.

A domain tree is simply constructed by labeling and characterizing each leaf organism of the phylogenetic tree by the type and number of protein domains in its genome. Although the hierarchical structure of the tree is identical, every domain has its

own correlated domain tree that depicts its unique distribution on the tree and its distinct evolutionary history.

The procedure for building domain trees is same as building domain combination trees, domain module trees, etc., the only difference being the evolutionary unit under consideration.

Domain Combinations

The domain assignment by the SUPERFAMILY database provides the position and length of each domain within proteins, therefore, for every protein, its composite domains and the domain orders in sequence are readily available (Table 5.1).

Table 5.1 Three examples of domain combinations. The boundary and identity of each domain are listed.

#	Protein	Dom 1	Dom 2	Dom 3	Dom 4
1	gi|49483312|ref|YP_040536.1|				
Position	271	19-128	130-254		
Identity	succinate dehydrogenase	d.15.4.2	a.1.2.1		
2	gi|15605900|ref|NP_213277.1|				
Position	253	1-62	63-143	144-253	
Identity	heterodisulfide reductase subunit C	N/A	a.1.2.1	N/A	
3	gi|49183540|ref|YP_026792.1|				
Position	380	70-147	148-250	251-313	47-69,314-375
Identity	iron-sulfur cluster-binding protein, putative	N/A	a.1.2.1	N/A	a.118.1.16

For a multi-domain protein, because of the lack of homologous domains with 3D structure, sometimes only part of the protein can be identified, leaving gaps in unassigned parts. These unassigned domains introduce complicated issues for the designation of domain combinations. First of all, gaps must be taken into consideration because the domain combination must reflect the real domain architecture of the protein. Second, gaps between domains may actually be linker regions that are not the actual structurally distinct domains, especially if the gaps are short. Given a histogram of domain length distribution (Figure 5.3), the lower bound cutoff of unassigned domains is selected to be 50 amino acids, covering 95% of all potential domains. The unassigned parts are labeled by "N/A" in the domain combination nomenclature (Table 5.1). Third, sometimes an unassigned part is as long as a few hundred or even over one thousand amino acids, so the unassigned part may contain multiple domains. But in our domain combination designation such a distinction is not possible, and there is no upper bound for gaps. If an unassigned part is longer than 50 aa, it is considered as one unknown domain. Similarly, unknown domains in the same domain architecture may be different. For instance, two proteins have the same domain combination, DomA N/A DomB, but the middle domain could be DomC or DomD. Their identity can only be predicted by sequence comparison. Finally, some structural domains consist of discontinuous sequence segments, and the HMM domain prediction algorithm is claimed to be able to pick up those cases. However, the prediction power of discontinuous domains has not been fully evaluated. In fact, the segmentation of domains make domain combination designation complicated; errors in prediction will make it even more controversial. In these cases, only the major fragment

of domains are considered; other parts are still viewed as part of the domain, but their existence is not labeled in the domain combination (Table 5.1, #3).

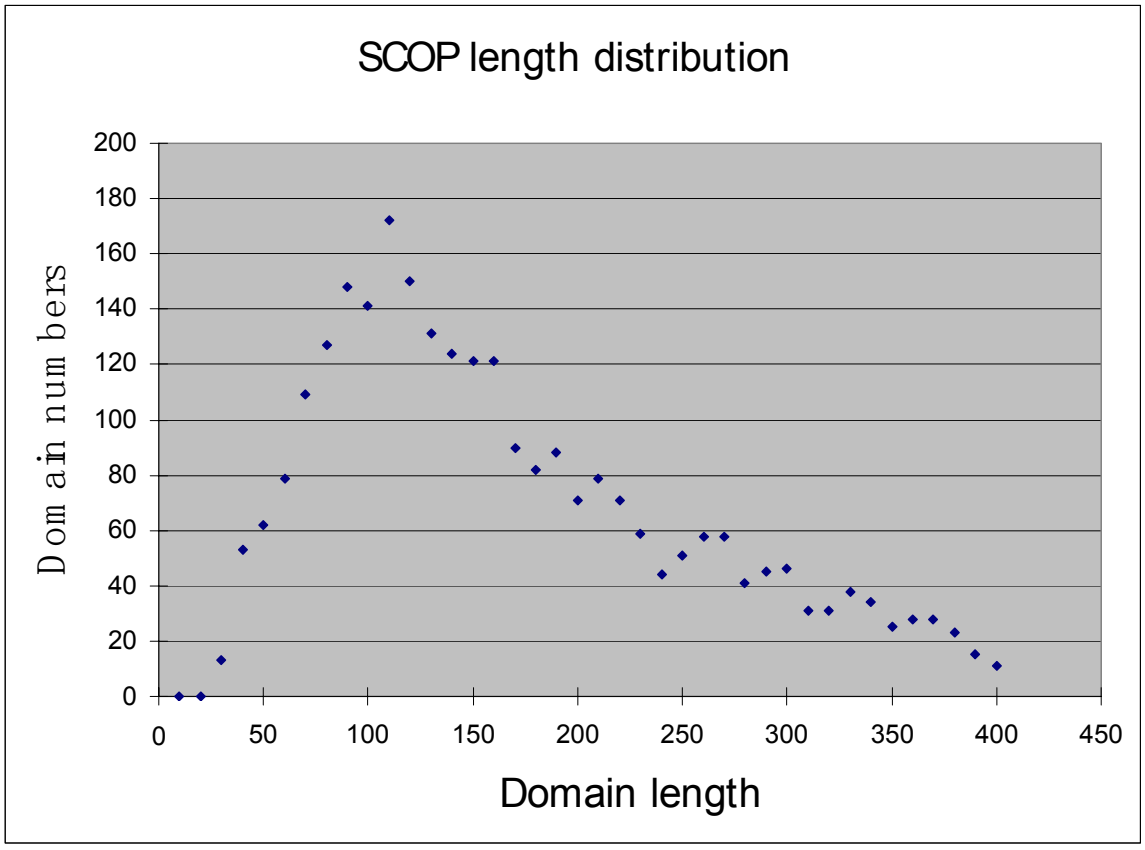


Figure 5.3 Protein domain length distribution according to SCOP 1.69. Domain definitions are retrieved from the ASTRAL database with 95% sequence identity.

Calibration of Domain Combinations

Identification of domain combinations is the first step towards an understanding of the evolutionary process of domain combination. However, the designation of domain combinations is imperfect due to our incomplete knowledge. In fact, there are systematic errors in domain combination identification: sometimes the predicted domain

combinations for two homologous proteins are different. Consider the case in Table 5.2, two dehydrogenases from *Bfra2* and *Hpyl1* are homologous, as can be seen in Figure 5.4, where the two sequences are fully aligned without major gaps by the Blastp program. The predicted domain architectures of the two proteins are similar except for a 66 aa gap in the first protein, which is labeled as an unknown domain according to our criterion. This gap region in another protein is predicted to be within the d.58.32.1 (Vanillyl-alcohol oxidase-like) domain. Further examination of the domain sequence and comparison with the SCOP definition confirms that the "gap" is an error and should belong to d.58.32.1.

Table 5.2 Comparison of domain combination assignment for two homologous proteins (D-lactate dehydrogenase) in *Bfra2* and *Hpyl1*.

Species	*Bfra2*		*Hpyl1*	
Gene ID	gi|53711720|ref|YP_097712.1|		gi|15645836|ref|NP_208014.1|	
Annotation	putative D-lactate dehydrogenase		D-lactate dehydrogenase (dld)	
	Scop id	region	Scop id	region
Domain 1	d.145.1.1	7-269	d.145.1.1	19-268
Domain 2	N/A	270-336		
Domain 3	d.58.32.1	337-514	d.58.32.1	243-527
Domain 4	a.1.2.1	491-631	a.1.2.1	498-628
Domain 5	N/A	632-956	N/A	629-948

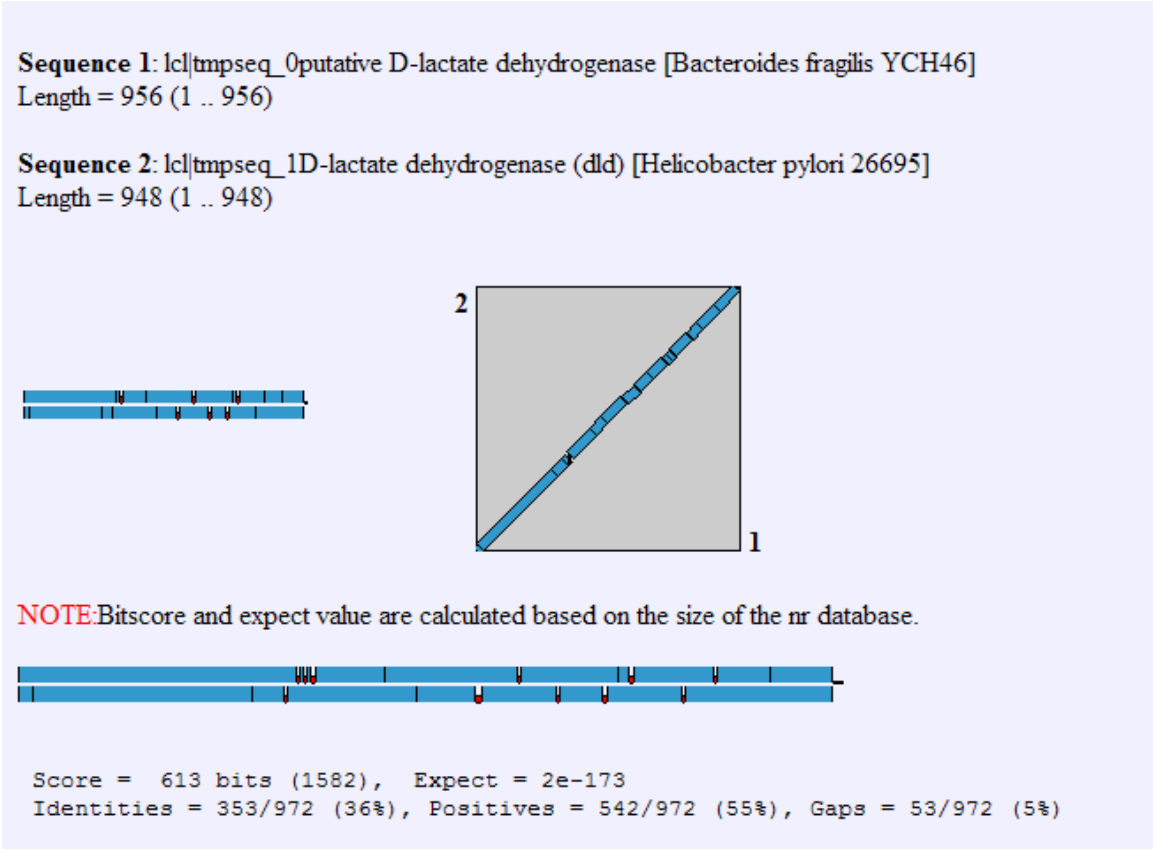


Figure 5.4 Sequence comparison of two lactate dehydrogenases using the Blastp program confirms that they are homologous proteins. The analysis uses the NCBI Blast 2 sequence server, which enables the pair wise sequence comparison by Blast.

This example represents a common problem in domain combination designation, which results from errors in domain boundary prediction. The junctions between domains are normally loop regions, so domain boundaries are not clear-cut. Sometimes, different structure classification schemes disagree on the assignment of domain boundary, even with a known protein 3D structure [28]. Therefore, variations in predicted domain

boundaries are inevitable. The errors in domain boundaries will be amplified when the unassigned gaps are taken into consideration, leading to the systematic errors outlined above.

One way to avoid this problem would be to first identity homologous protein families that can be aligned across their entire length, and then use HMM domain assignment algorithms to assign domains to these protein families. This procedure ensures that homologous proteins always have the same domain architecture in our prediction. However, this method requires the recalculation and regeneration of the entire SUPERFAMILY database, which is time consuming. We need an alternative method to correct the errors in our current data.

The example in Table 5.2 indicates that due to the errors in domain boundary prediction, homologous proteins are mistakenly assigned different domain architectures. In order to correct for these, domain combinations that might actually be associated with the same architecture should be identified. We developed a procedure to group together similar domain combinations based on two criteria: similarity in domain combination and closeness in total protein length. The first criterion mainly handles combinations that differ in one or more unassigned parts, similar to the case in Table 5.2. Putative unassigned domains could acturally be loop regions, one or more unknown domains, or even familiar domains that were missed by the domain prediction algorithms. The different possibilities arising from unassigned parts are scanned by their total protein length. Similarity in protein length should be statistically significant, assuming the normal distribution in protein lengths. The program iteratively finds and compares

potential similar domain combinations and clusters those to satisfy the two criteria for calibrated domain combinations. Many domains have over one hundred or a few thousand combinations and some combinations contain up to 13 unassigned parts, so the calibration of domain combinations is a time consuming step.

Although the procedure can by no means fully correct the errors due to inaccurate domain assignments, the result is quite significant. By a rough estimate, the majority of errors in domain combination can be recovered by this procedure without expensive genomic sequence comparison, such as the case in Table 5.2. In addition, because similar combinations are grouped together, the side effect of the calibration step is reduction in the total number of combinations. Initially, the total number of combination was 51204, which dropped to 33513 after the calibration step.

PDB-validated Tree

There are errors at many steps in the data preparation, including the domain combination error we just discussed. In fact, while correcting systematic errors our calibration step introduces its own false positives even with the careful selection of parameters. These errors can cause serious problems when investigating the evolutionary history of each individual domain or domain combination. One mistake may cause the switch of a domain's present/absent state in an organism, which will further alter our understanding of its origin, loss or transfer and other evolutionary events.

The PDB-validated tree is designed to identify errors within a domain or in domain combination assignments in the tree of life. Because the SCOP classification is based on proteins with "real" 3D experimental structures in the PDB, these "true" domains exists within the genome of the species from which they were extracted. A PDB-validated tree incorporates the domain assignment in all the complete genomes in our data set, as well as the PDB-confirmed domain in certain species, which may or may not exist in our complete genomes set. The purpose of the PDB-validated tree is to validate the prediction of domains in other species using the existence of PDB-confirmed domains in certain species as the positive control. For example, domain A in the PDB is found in *Equus caballus* (horse), and *Equus caballus* is not in our complete genome set. The PDB-validated domain tree of A includes all the over 300 species as well as horse. The true assignment in horse can help confirm the accuracy of the prediction of domain assignment in other species, when they are closely related to horse. Using this method over 90% of the domain trees can be confirmed with confidence.

5.2.2 Results

Divergence of domain families

Different domain families within a superfamily originated from a common ancestor, but their sequences have diverged so much that their evolutionary relationship could only be recognized by structural and functional relatedness. The distribution of

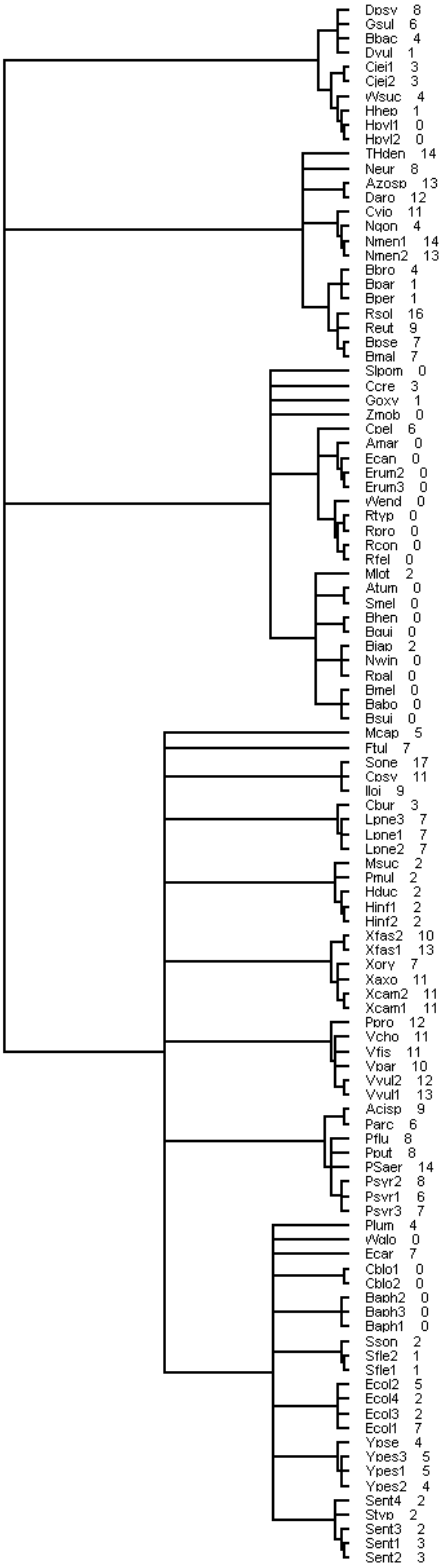
different families within the same superfamily indicates where the divergence events happened in the tree of life.

Pilin refers to a class of fibrous proteins that oligomerize and form the pilus structure in many bacterial species. Bacterial pili are involved in adhesion to surfaces and conjugate with other bacteria. The pilin superfamily (d.24.1) in SCOP is represented by two families, pilin (d.24.1.1) and TcpA-like pilin (d.24.1.2), the later one being the toxin-coregulated pilus from *Vibrio cholerae*. As shown in the domain trees of the two families (Figure 5.5), the pilin family is found in many bacteria species but not in Archaea and Eukaryotes, so it probably originated in the common ancestor of all bacterial organisms. The TcpA-like pilin family is only found in two species, *Vibrio cholerae* and *Vibrio fischeri*, but not in other bacterial species, so it was probably diverged after duplication from the pilin family in one of the ancestor *Vibrio* species.

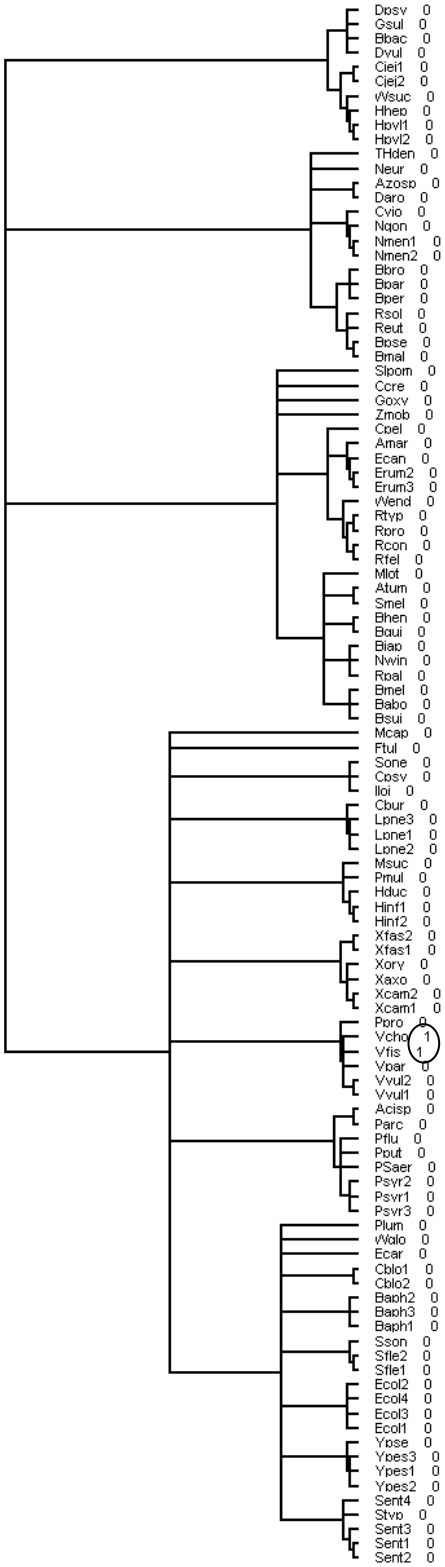
The domain trees explicitly illustrate the evolutionary history of the pilin superfamily (d.24.1), including the origin of the pilin protein family (d.24.1.1) in the root of bacteria, domain duplication in each genome, gene loss in some bacteria clades, and particularly the formation of a new family (d.24.1.2) in the *Vibrio* species.

Figure 5.5 The domain trees of the pilin family (d.42.1.1) and the TcpA-like family (d.42.1.2). Both families exist in bacteria exclusively, and only the proteobacteria taxa within the bacteria are shown due to lack of space. The number behind each species represents the abundance of the domain family in the genome of the species. The d.42.1.1 family exists in most of the proteobacteria species, except in several α proteobacteria and γ proteobacteria clades that might result from gene loss. The d.42.1.2 family only exists in Vcho and Vfis, which are marked in the figure.

d.42.1.1



d.42.1.2



δ/ε Proteo-bacteria

β Proteo-bacteria

α Proteo-bacteria

γ Proteo-bacteria

Since the pilin proteins are single domain proteins, its evolutionary history can be accurately represented by its domain tree. For protein domains that form various proteins with different domain architectures, domain combination trees are more informative, but more prone to error.

Evolution of Domain Combinations

Recombination of existing domains is a typical way of forming novel proteins. The domain combination trees can illustrate the domain recombination processes in the tree of life. For instance, the a.109.1.1 domain, named by SCOP as the Class II MHC-associated invariant chain ectoplasmic trimerization domain, exists in two domain architectures. These differ at the C-terminus by one g.28.1.1 domain, the thyroglobulin type-1 domain. The two domain architectures are two isoforms of the CD74 antigen protein. The domain combination tree of a.109.1.1 shows that the isoform one protein, N/A ~ a.109.1.1, originated from the ancestor of *Amniota*, and the isoform two protein, N/A ~ a.109.1.1 ~ g.28.1.1, appeared in *Euteleostomi* by appending a thyroglobulin type-1 domain at the C-terminus of isoform one (Figure 5.6).

Some domains only combine with one other domain and exist in one type of domain combination, e.g., the N- and C- terminal subunit of aldehyde ferredoxin oxidoreductase (SCOP id d.152.1.1 and a.110.1.1). In this case, the entire combination evolves as one unit, and the combination tree is the same as the domain tree.

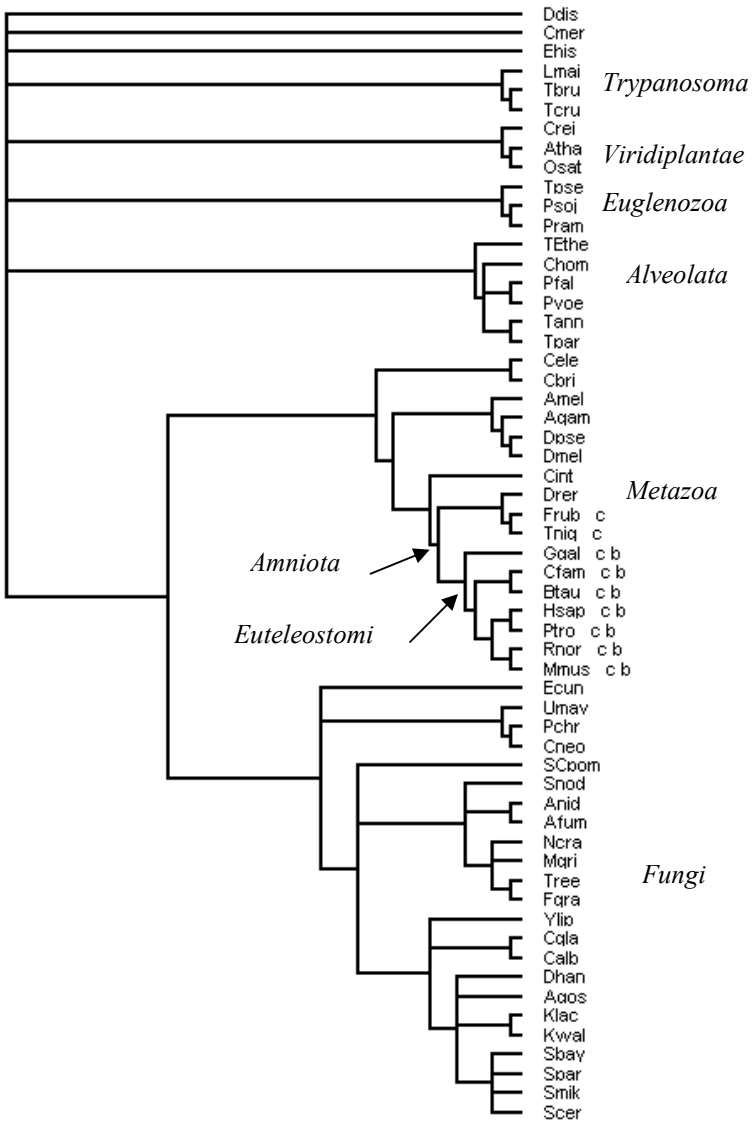


Figure 5.6 The domain combination Eukaryotic tree of a.109.1.1, the Class II MHC-associated invariant chain ectoplasmic trimerization domain. The letters next to the species represent different combination types. In this case, type b corresponds to N/A ~ a.109.1.1 and c represents N/A ~ a.109.1.1 ~ g.28.1.1, where N/A is an unknown domain.

Divergence of Domain Combinations

Generally, during evolution the divergence and recombination of protein domains took place concurrently resulting in complex domain architectures. After an ancestor domain diverged into different families, the descendant domain families can combine with different domains and form different domain architectures. This combination after divergence is the typical scenario. Since different families have little sequence similarity and distinct function, their domain partners are usually different.

On the other hand, within an existing domain combination, one or more constitute domains can diverge into distantly related families, while still keeping the original combination. This scenario, called the divergence of constant domain combinations, is found in the evolution of over thirty SCOP superfamilies using domain combination trees.

For example, Clathrin adaptor appendage domain superfamily (b.1.10) is a 120 aa all-beta domain that contains three families. Each family and the combination types it forms are listed in Table 5.3, their 3D structures are compared in Figure 5.7.

All three families combine with the Clathrin adaptor core protein a.118.1.10 at the N-terminus (type A in Table 5.3), which infers that the a.118.1.10 ~ b.1.10 combination remains even after the b.1.10 superfamily diverges into three different families. Furthermore, in type B of b.1.10.1 and b.1.10.3, an additional appendage domain from different families of the same superfamily is added to the C-terminal of the a.118.1.10 ~ b.1.10 combination. In this case, both domains in the combination diverged into different families.

Table 5.3 The domain combinations of three clathrin adaptor appendage domain and their evolutionary origin in the tree of life.

Scopid	Scop Name	Abbr*	Combination	Origin
b.1.10.1	Alpha-adaptin ear subdomain-like	A	a.118.1.10 b.1.10.1	Eukaryote
		B	a.118.1.10 b.1.10.1 d.105.1.1	
		C	b.1.10.1 d.105.1.1	
b.1.10.2	gamma-adaptin C-terminal appendage domain-like	A	a.118.1.10 b.1.10.2	Eukaryote
		B	a.118.9.2 a.7.8.1 b.1.10.2	Fungi/Metazoa group
		C	a.7.8.1 b.1.10.2	
b.1.10.3	Coatomer appendage domain	A	a.118.1.10 b.1.10.3	Eukaryote
		B	a.118.1.10 b.1.10.3 d.105.1.2	
		C	a.118.1.2 b.1.10.3	
		D	b.1.10.3 d.105.1.2	

a.118.1.10: Clathrin adaptor core protein

d.105.1: Subdomain of clathrin and coatomer appendage domain

d.105.1.1: Clathrin adaptor appendage, alpha and beta chain-specific domain

d.105.1.2: Coatomer appendage domain

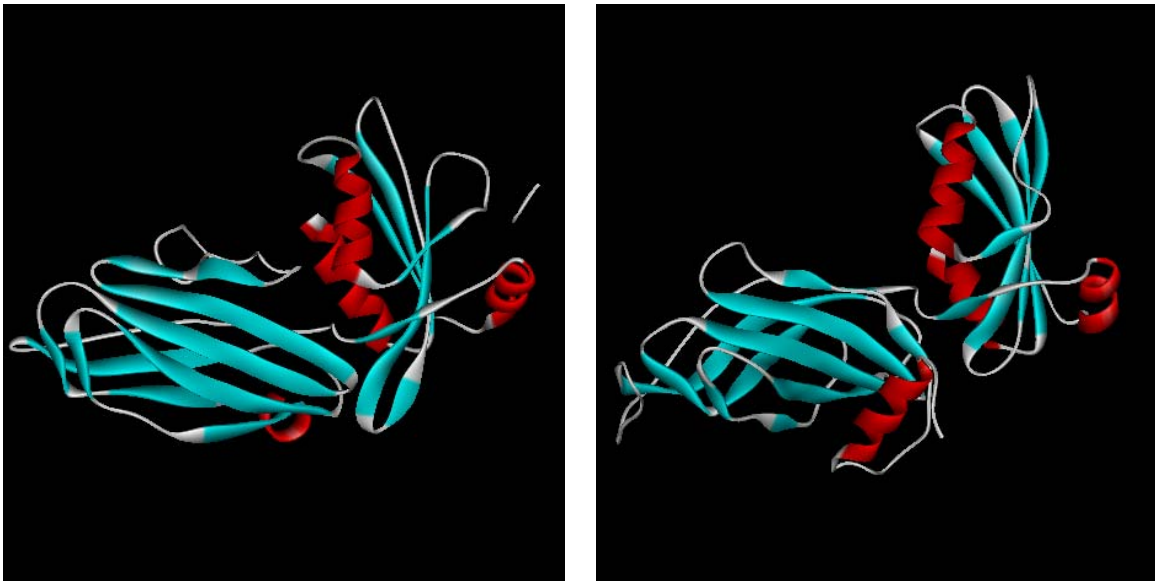
a.118.9.2: VHS domain

a.7.8.1: GAT domain

*** the letters represent domain combinations**

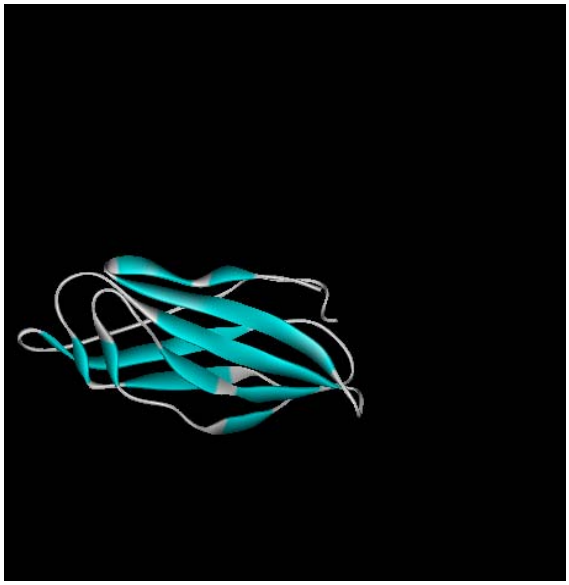
The domain combination trees (Figure 5.8) suggest that both the formation of the a.118.1.10 ~ b.1.10.* ~ d.105.1.* combination and divergence of the combination into three families originated from the common ancestor of Eukaryotes. Alpha/beta-adptin (with b.1.10.1) and gamma adaptin (with b.1.10.2) are functionally more related, while the domain architecture of coatomer (with b.1.10.3) is more similar to that of alpha/beta-adptin. As a consequence, either alpha/beta-adptin appeared first and developed into gamma adaptin and coatomer, or one of the other two proteins came earlier and first evolved into alpha/beta-adptin. It's worth noting that the d.105.1 appendage domain superfamily only exists in these combinations.

At the root of the Fungi/Metazoa group, b.1.10.2 combined with a.118.9.2 ~ a.7.8.1 and formed a new combination that did not exist in the other two clathrin adaptor appendage domain families.



b.1.10.1 ~ d.105.1.1 (1kyf)

b.1.10.3 ~ d.105.1.2 (1pzd)



b.1.10.2 (1gyv)

Figure 5.7 Comparison of the 3D structures of the remotely related domain combinations that contain the same domain architecture at the superfamily level but differ at the family level. The b.1.10 domains on the left contain two parallel beta-sheets, and the d.105.1containing one beta sheet locates on the right.

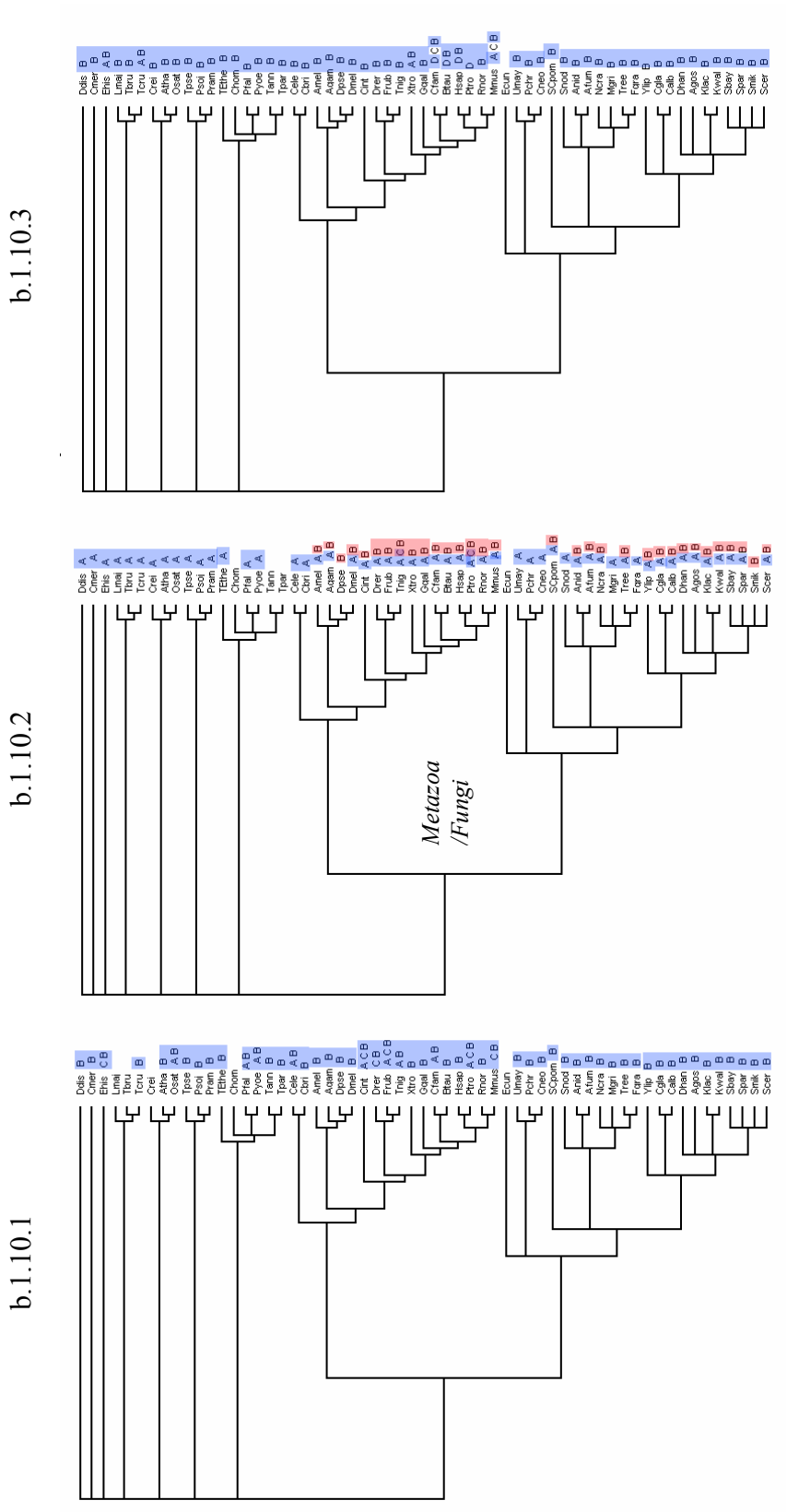


Figure 5.8 The domain combination Eukaryotic trees of three clathrin adaptor appendage domain families.

Blue represents the a.118.1.10 ~ b.1.10.3 ~ d.105.1.2 combination and its derivatives;

Functionally Related Domain Combinations

The PDB-validated domain combination trees of d.58.31.2 ~ a.89.1.1 and d.58.31.1 are shown in Figure 5.9.

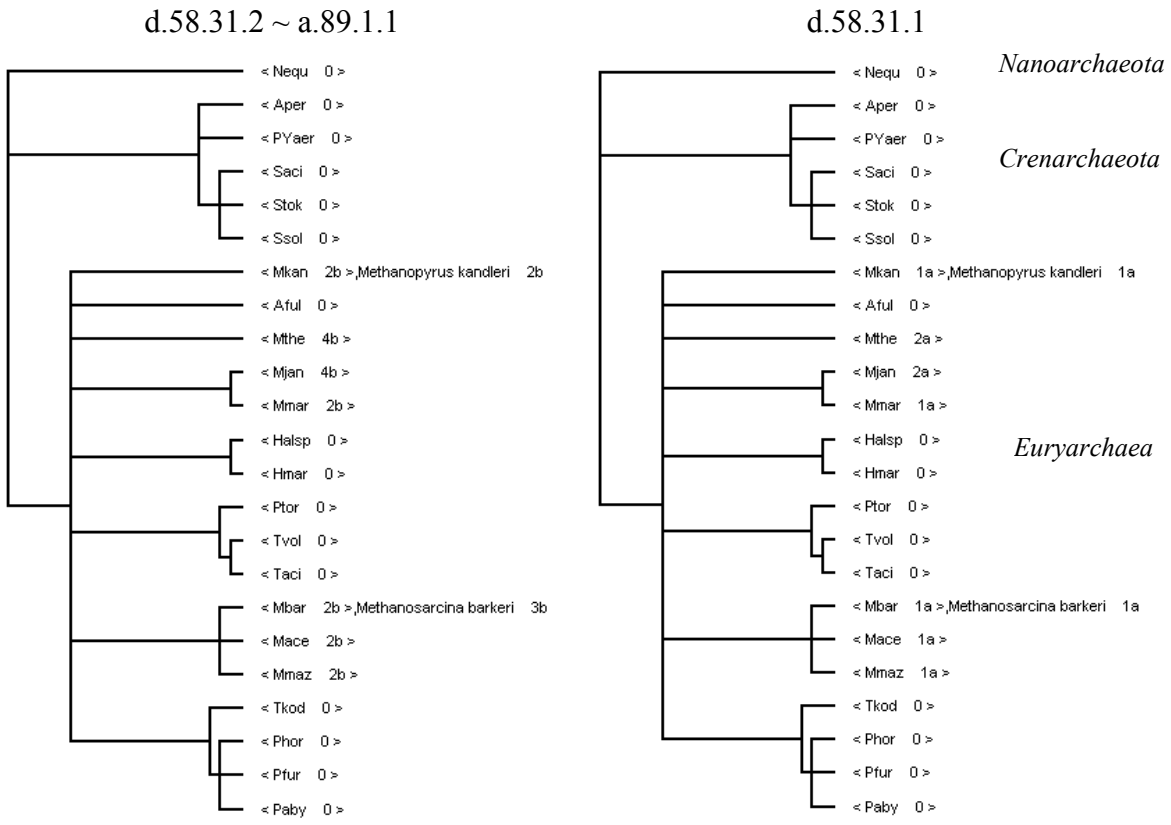


Figure 5.9 The PDB-validated archaeal trees of two domain combination, d.58.31.2 ~ a.89.1.1 and d.58.31.1, are compared. The numbers and types in the parenthesis, such as '<Mkan 2b>', represent the predicted domain combinations; those outside of the parenthesis, such as 'Methanopyrus kandleri 2b', represent the PDB containing combinations that have solved 3D structures from that species.

Figure 5.9 shows that the distributions of two combinations in the archaea tree are identical; and the abundance of the d.58.31.2 ~ a.89.1.1 combination in any species is

always twice that of d.58.31.1. The correlation in the distribution of the two proteins occurs because the two proteins are part of the same protein complex.

Methyl-coenzyme M reductase, found exclusively in archaea, consists of three polypeptide chains. α- and β- chains are homologous, each contains an N- terminal (d.58.31.2) and a C- terminal domain (a.89.1.1). γ- chain is a single domain protein (d.58.31.1), which is remotely related to the N-terminal domain of the α- and β- chain.

In this case, the structure and function of methyl-coenzyme M reductase is already known. In general, the correlation of the distribution of domains or domain combinations of unknown function can potentially be used to infer their functional information.

5.2.3 Discussion

Taxonomy-based Phylogeny

The major problem with the taxonomy-based phylogeny is that it is not a binary tree, which the phylogenetic tree of life intrinsically should be. The six to seven major hierarchical levels of the taxonomy lead to multiple clades at the same level whose evolutionary relationships are not resolved. As a result, the evolutionary origin of a domain or domain combination determined by the taxonomy-based phylogeny is biased towards the higher level of the phylogeny. For instance, according to the phylogeny in Figure 5.9, the origin of the d.58.31.2 ~ a.89.1.1 is the root of *Euryarchaea*, accordingly, many clades such as *Thermoplasma* and *Halobacteria* must have lost this protein during evolution. In fact, this protein is involved in the formation of methane, a function that was only developed in methanogenic archaea relatively late in evolution. Therefore, if the evolutionary relationship of the seven classes within the *Euryarchaea* kingdom is resolved, the origin of this protein is more likely to be found in the root of all methanogenic archaea, and hence no gene losses in multiple lineages are necessary. Clearly, the bias of evolutionary origin is also companioned with the over-estimation of the extent of gene loss.

The problem brought about by taxonomy can be corrected by using the bifurcating phylogenetic trees that contains detailed evolutionary relationship of every taxa. Currently, some branching of the tree of life are still unsolved and in debate, such as the divergence of the last common ancestor, and the partition of animal, plant and fungi.

In those cases, the taxonomy-like phylogeny that allows multiple leaves under a node have to be used.

Horizontal Gene Transfer

It is widely agreed that horizontal gene transfer (HGT) is a major force in the evolution of prokaryotes [29-32]. Within-species genome comparisons suggested that up to 20-30% of genome variation is due to this process [33, 34]. Horizontal gene transfer is so pervasive that some have claimed that the reconstruction of a universal tree of life is not possible [35]. Our approach assumes the existence of an organismal phylogeny that represents the entire history of life, but HGT is still a critical issue to be considered, especially in studying the evolutionary history of individual domain or domain combinations.

Because the multiple sequence alignment by HMM algorithm does not distinguish orthorlogs and paralogs, and after domain assignment, only the presence and absence of domains are evaluated in domain trees and domain combination trees, horizontal gene transfers of homologous proteins within closely-related species are not considered in our analysis. Although HGT might happen extensively among prokaryote species, since they did not change the domain content of species, nor did they affect the identity of each individual domain, these types of HGTs are ignored in our analysis. Only foreign non-homologous proteins that are transferred from distantly-related species can be recognized by domain trees, so this type of HGT should be considered.

A possible HGT can be inferred if the distribution of a domain in the domain tree forms more than one distinct phylogenetic group. For example, phycocyanin-like phycobilisome proteins (a.1.1.3) are light harvesting antennae of photosystem II. Its domain tree (Figure 5.10) shows that it only exists in cyanobacteria and red algae, which strongly supports the origin of this domain in red algae by acquisition from endosymbiotic cyanobacteria.

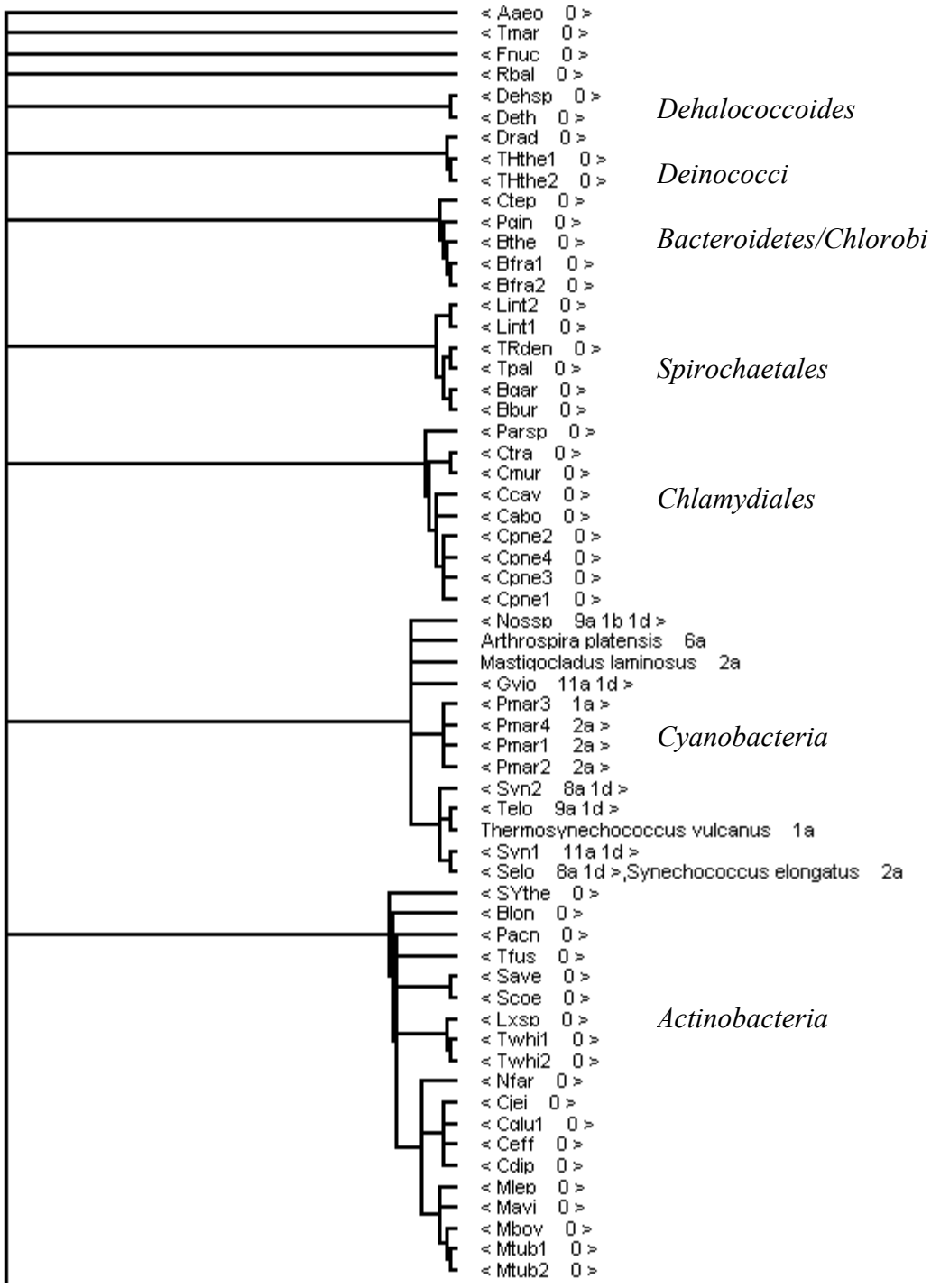


Figure 5.10 The PDB-validated domain combination trees of phycocyanin-like phycobilisome proteins (a.1.1.3). (A) Part of the bacteria tree, a.1.1.3 exists only in cyanobactria species.

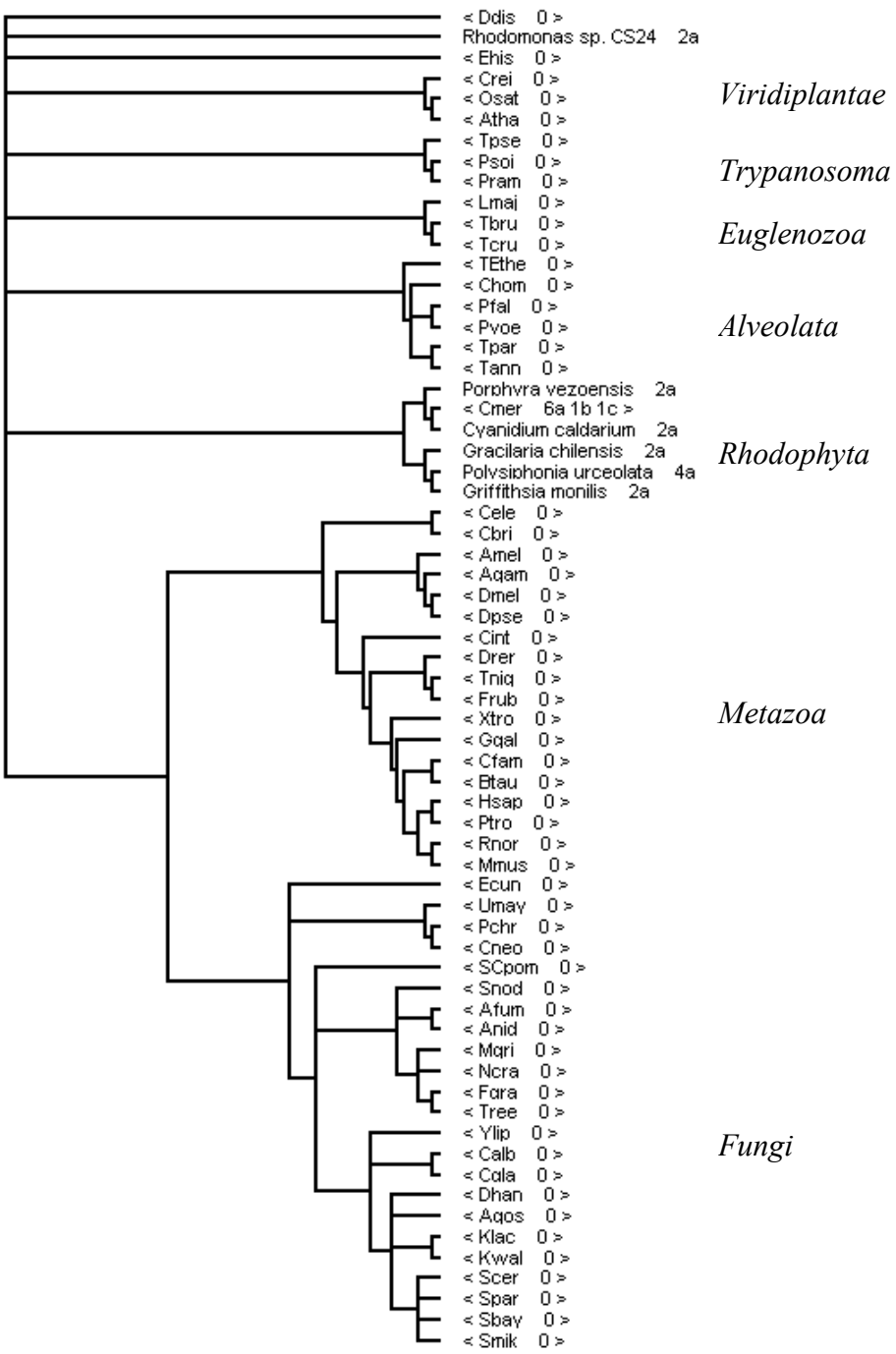


Figure 5.10 continued (B) In the eukaryote tree, a.1.1.3 only appears in one red algae. PDB contains a.1.1.3 in many red algae species.

The same domain distribution can also be explained by gene loss or convergent evolution. If the two distinct phylogenetic groups are far away from each other on the tree, massive gene losses are needed to account for the domain distribution. In the a.1.1.3 example, there should be about 20 gene losses in most of the major kingdoms within bacteria and eukaryote. Therefore, HGT is a more favorable explanation.

Convergent evolution is the process whereby organisms not closely related, independently evolve similar traits as a result of having to adapt to similar environments. Convergent evolution of domains means *de novo* domain genesis in different species, although this might exist, is very rare, and is hard to detect and confirm. However, convergent evolution of domain combinations, which only involves forming the same domain connection independently in different lineages, is relatively easier, if every domain in the combination already exists in the genome.

5.3 Systematically Derive the Origin and Evolution of Domains

As shown those far, the evolutionary history of a domain can be readily viewed in its domain tree and domain combination tree. However, the information in those trees can only be inferred by manual analysis. As shown, by visual inspection of the distribution of domains in the leaf species and the hierarchy of the tree, researchers can find the evolutionary origin of domains, as well as domain loss and transfer events.

However, there are about 2,800 protein domains at the family level and over 30,000 combination types after calibration. To examine the phylogenetic tree for each domain and domain combination is tedious and time-consuming. Traditionally scientists interested in one specific protein or domain need only know the evolutionary history of one protein/domain. In the era of complete genomes and comparative genomics, researchers are more interested in the evolution of the protein universe and the relationship of multiple domains during evolution. Indeed, even if one can possibly look at every tree and get the evolutionary history for each domain, the comparison of the evolution of different domains is still unmanageable, because the number of pairs for comparison is much larger than that of single domain. Moreover, comparison of the history of domain combinations, which involve examination and comparison of the history of several domains, makes it even more complicated.

In fact, with the phylogenetic information embedded in the trees, there is a need to develop a procedure to automatically retrieve this information from the trees without human intervention.

5.3.1 Method

The question is, given a phylogenetic tree and the distribution of a domain or domain combination among all the leaf species, how can we find its evolutionary origin and the present/absent status of the domain (combination) in all the ancestor species in the tree. That is, we want to reconstruct the protein domain content of ancestor organisms in the tree of life. This approach will explicitly show which domains and how many of them were present in the ancestral genomes and how the domain content of ancestral and present day genomes has been shaped by the processes of domain loss, domain duplication, HGT, and domain genesis.

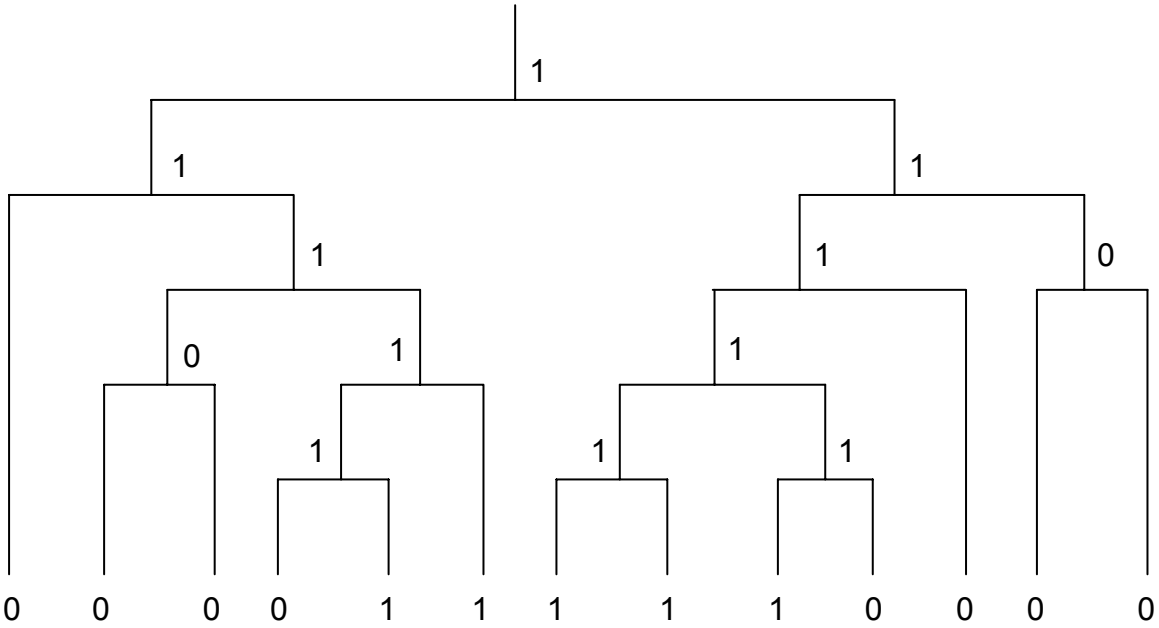
Assume we have a phylogenetic tree and the present day presence pattern of a certain domain, as shown in Figure 5.11. Here, we only consider the presence and absence of domains, which are represented by 1 and 0, respectively. There are totally four evolutionary processes that change the presence status of domain content: namely, vertical inheritance, domain loss, horizontal gene transfer and domain genesis. (Domain duplication and recombination do not affect domain content.) We assign each process a penalty score according to their relative ease of occurring during evolution, one example is listed in Table 5.4.

Table 5.4 Penalty scores for evolutionary processes.

Vertical Inheritance	0
Domain Loss	1
HGT	10
Domain Genesis	10

The score for vertical inheritance is set to be 0, because it is the default process that ensures carrying of genetic information from parents to descendants. In general, HGTs are evolutionarily more stringent than domain loss, so their penalty score is much higher than that of domain loss. As an example, in Table 5.4 the scores for HGT and gene loss are set to be 10 and 1 respectively. Domain genesis, the *de novo* creation of domains, took place only once for each domain during evolution, if convergent evolution is omitted; where HGT apparently can cause multiple origins for a domain. If a domain trees shows n origins, one belongs to domain genesis, the other n-1 is accounted by HGT. Since, both domain genesis and HGT result in one origin in the tree, it will be convenient to set the penalty score for HGT and domain genesis to be equal.

A: Without HGT



B: With HGT

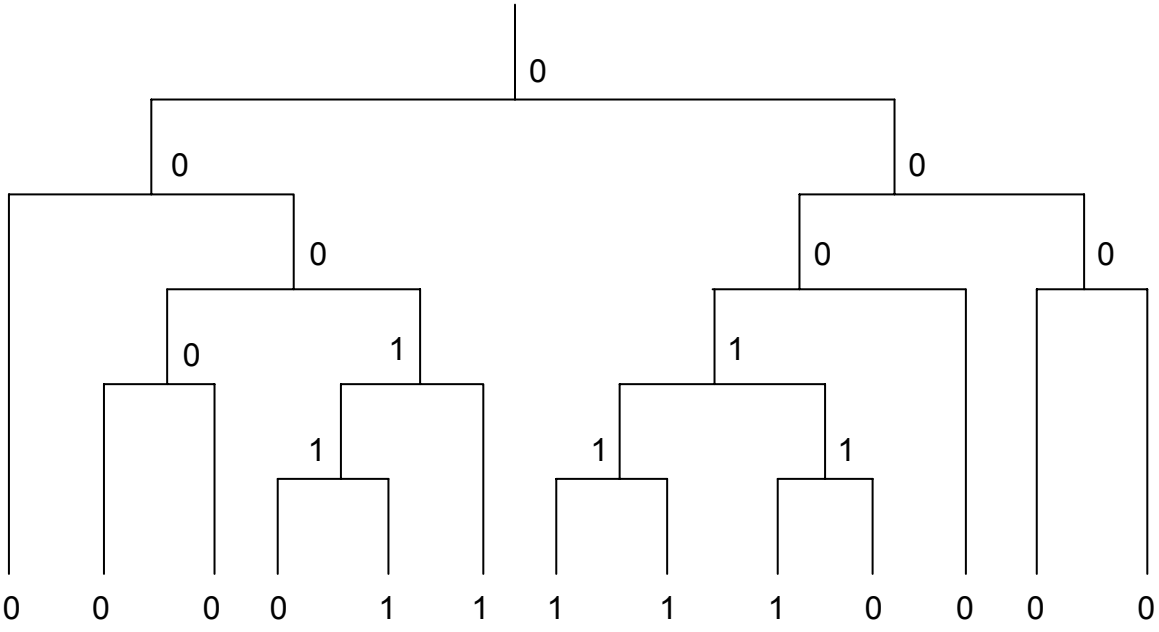


Figure 5.11 Two possible evolutionary histories for domains that result in the same current domain distribution

To find the ancestor domain content that best fit the current domain distribution is equivalent to finding the most parsimonious present/absent dataset for each node on the tree, in order to minimize the total score for the whole tree.

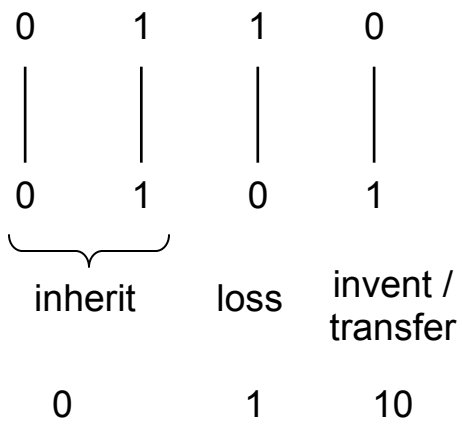


Figure 5.12 The four possible evolutionary events between a pair of parents/children.

Each node has two possible conditions, presence or absence. There are four possible combinations between a pair of parents and children, as illustrated in Figure 5.12. The most straight forward method to find the most parsimony dataset would be to traverse the whole tree and test every possible combination. But the calculation complexity is 2^N, N being the total number of nodes on the tree. The exponential complexity of this algorithm is not applicable when the number of species in the tree is large.

Instead, we developed an iterative algorithm that at each step uses the relative penalty score to invent or transfer as a cutoff to simplify calculation. The time complexity of this algorithm is estimated to be $N^2 \log N$.

5.3.2 Results

The Statistics of the Protein Domain Universe

SCOP 1.69 classified proteins with known 3D structure into 945 folds, 1539 superfamilies and 2845 domain families. The total number of domain combinations formed by the 2845 families in the species with complete genomes (Table 5.5) are more than 10 fold the number of domain families and will increase as more genomes are sequenced.

Table 5.5 The total types of domain combinations. The second and third rows list the number of domain combination before and after correction for errors. The fourth row lists combinations with high abundance. Specifically, for a certain domain, only the combinations that cover the top 95% of domain abundance are counted.

Domain Families	2845
Initial Domain Combinations	51204
Domain Combination After Calibration	33513
Domain Combination 95% abundance	15532

The abundance of combinations is unbalanced, with only half of the combination types account for 95% of total combinations. Many combinations (proteins) are species specific and only exist in one or limited organisms, while some combinations, which originated in the last common ancestor and duplicated multiple times during evolution, are very abundant. In general, the abundance of domain combinations follows the power law (Figure 5.13). Among the top 20 abundant domain combinations, many are single domain proteins, and some only exists in Eukaryotes, such as the rhodopsin-like domain (Table 5.6). Protein kinase domain, ABC transporter ATPase domain and the zinc-binding domain are among the most abundant domains either by themselves or combine with other domains.

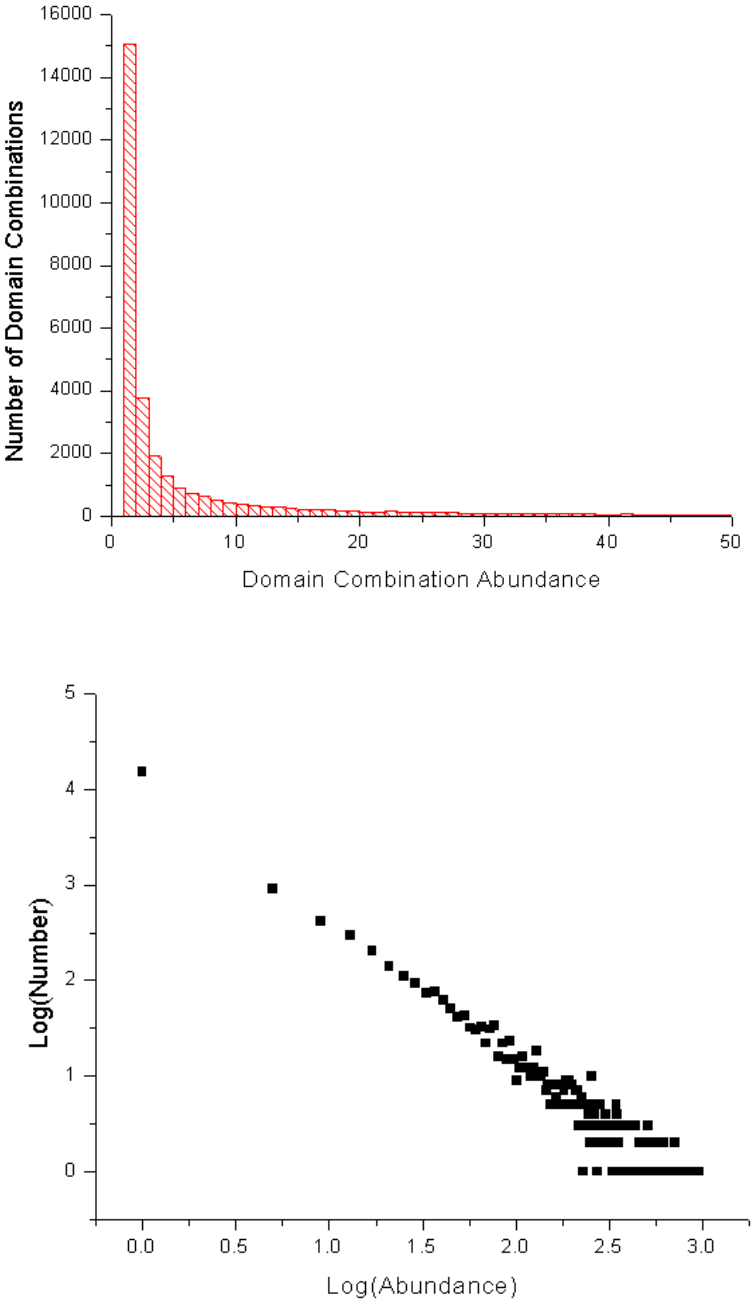


Figure 5.13 The histogram and the double-log plot of domain combination distribution in species with complete genomes shows a power-law distribution.

Table 5.6 The top 20 most abundant domain combinations in all.

Abundance	Domain Combination	Protein
12121	f.13.1.2	Rhodopsin-like
9730	c.2.1.2	Tyrosine-dependent oxidoreductases
7982	a.212.1.1 N/A g.37.1.1 g.37.1.1 g.37.1.1 g.37.1.1 g.37.1.1 g.37.1.1 g.37.1.1 g.37.1.1	zinc finger protein 195
6619	c.37.1.19 N/A N/A c.37.1.19 N/A	Tandem AAA-ATPase domain containing
6561	c.37.1.12	ABC transporter ATPase domain-like
6258	f.38.1.1	Glycerol-3-phosphate transporter
5598	c.94.1.1	Phosphate binding protein-like
5477	d.144.1.7	Protein kinases, catalytic subunit
4921	f.38.1.2	LacY-like proton/sugar symporter
4881	c.37.1.8	G proteins
4864	d.108.1.1	N-acetyl transferase, NAT
4264	d.144.1.7 N/A	Protein kinases related
3930	a.4.5.37 c.94.1.1	LysR family regulatory protein
3669	a.104.1.1	Cytochrome P450
3565	N/A d.144.1.7	Protein kinases related
3071	c.37.1.12 N/A	ABC transporter ATPase domain-like containing
2913	N/A d.144.1.7 N/A	Protein kinases related
2859	N/A c.37.1.12 N/A c.37.1.12 N/A	ABC transporter ATPase domain-like containing
2856	e.23.1.1	Acetyl-CoA synthetase-like
2834	b.35.1.2 c.2.1.1	zinc-binding dehydrogenase

Not only the total abundance of combinations, but also the type of domain combination that each domain can form is skewed and follows a power law (Figure 5.14). The top domains that can form more type of combinations usually have more adjacent domains (Table 5.7).

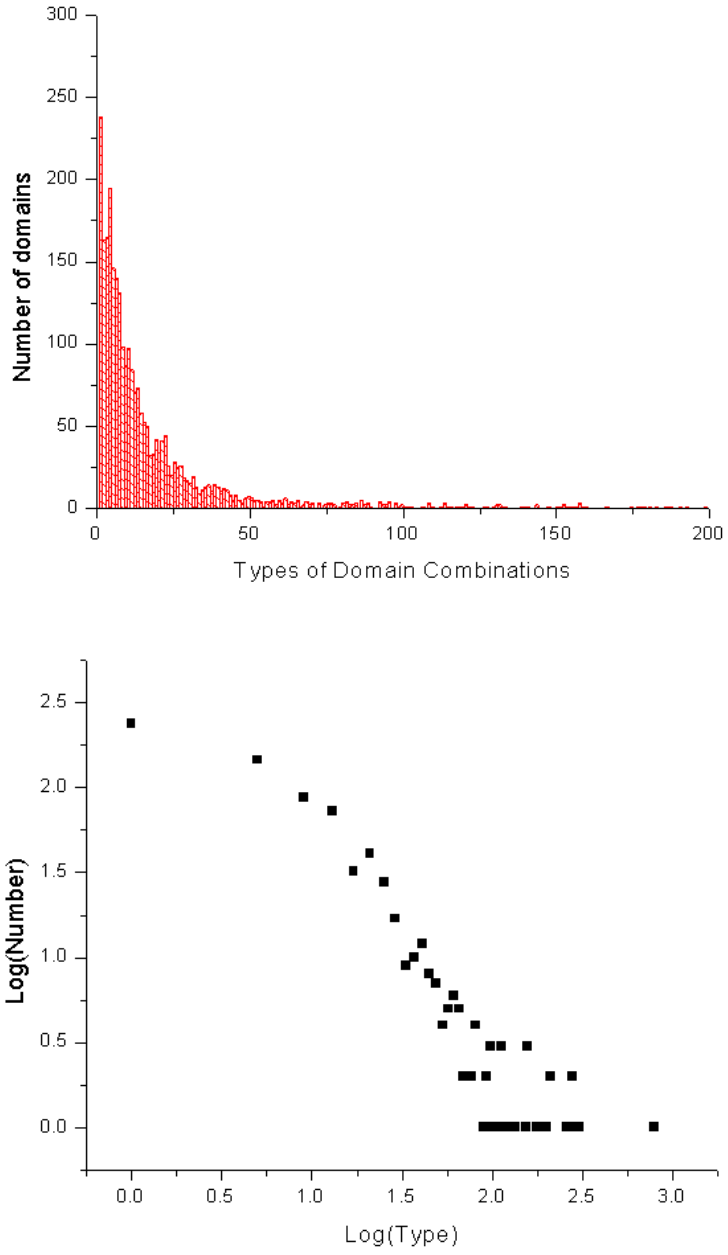


Figure 5.14 The histogram and the double-log plot of the number of combinations that each domain is involve in.

Table 5.7 The top 10 domains with most number of combination types.

	# combination	# Adjacent Domain	Domain Description
g.3.11.1	952	58	EGF-type module
d.144.1.7	785	31	Protein kinases, catalytic subunit
g.3.9.1	675	43	Growth factor receptor domain
d.211.1.1	464	48	Ankyrin repeat
a.118.8.1	452	37	Tetratricopeptide repeat (TPR)
b.55.1.1	431	49	Pleckstrin-homology domain (PH domain)
b.69.4.1	431	16	WD40-repeat
b.1.1.4	414	38	I set domains
b.1.2.1	383	41	Fibronectin type III

Above all, the analysis indicates that the types and abundance of domain combinations is large and unbalanced, their current distribution are due to their evolutionary history.

The General Trends of Protein Domain Evolution

Using our automatic procedure, we can derive the origin of all domains and domain combinations, as well as other evolutionary events, and illustrate those events on the tree. Table 5.8 lists the number of domains and domain combinations in the major lineages of the tree of life. 1592 domains are predicted to be in the root of the tree, accounting for more than half of the total domains. This prediction is significantly higher than what generally believed and the data by the Venn diagram in Chapter 3. Part of the reason is that the root of the tree is still not solved, so that any domains that were shared

by two superkingdoms are counted as evolved in the LUCA. Endosymbiosis of mitochondria and chloroplasts and horizontal gene transfer across superkingdoms are another reason. There are also a significant number of domains emerging at the root of bacteria and eukaryotes. Nevertheless, these data suggest that a large proportion of protein domains were invented in the root or after the separation of the three major superkingdoms but before the further differentiation of each lineage. When tracing down the tree from the root, the number of novel domains invented at each node decreases. Many branches and species don't invent any domains. Above all, during the approximately two billion years after the appearance of the fist eukaryotic cell, only 708 domains, less than 1/4 of the total domains, had been invented.

Conversely, the evolutions of domain combinations are different. There are 3015 combinations in the root of tree, which is quite large, but only account for 1/10 of the total combination. Relatively more combinations were invented in the descendant nodes of the tree, as proven by the ratio of number of combination and domain (Table 5.8). The data suggest that once protein domains have been generated and inherited in genomes, biological organisms tend to create new proteins and functions through recombination and duplication of existing domains, rather than creation of new domains *de novo*.

Table 5.8 The number of domain and domain combinations that originated from each node. The last column lists the ratio of combination / domain.

Nodes	# Domain	# Combination	# Cb / Dm
cellular organisms	1592	3015	1.89
Eukaryota	426	3579	8.40
Fungi/Metazoa group	27	339	12.56
Bilateria	74	628	8.49
Coelomata	25	758	30.32
Chordata	11	278	25.27
Euteleostomi	47	1123	23.89
Amniota	5	220	44.00
Eutheria	7	426	60.86
Euarchontoglires	2	66	33.00
Homo/Pan/Gorilla group	1	73	73.00
Homo sapiens	2	40	20.00
Fungi	7	189	27.00
Viridiplantae	2	71	35.50
Alveolata	0	13	
Bacteria	112	520	4.64
Firmicutes	15	114	7.60
Bacilli	11	64	5.82
Mollicutes	0	5	
Lactobacillales	4	21	5.25
Bacillales	8	42	5.25
Proteobacteria	67	452	6.75
Cyanobacteria	31	161	5.19
Alphaproteobacteria	3	58	19.33
Betaproteobacteria	6	66	11.00
Gammaproteobacteria	18	137	7.61
delta/epsilon subdivisions	1	7	7.00
Enterobacteriaceae	20	88	4.40
Actinomycetales	13	198	15.23
Archaea	7	74	10.57
Euryarchaeota	16	135	8.44
Thermoprotei	0	12	

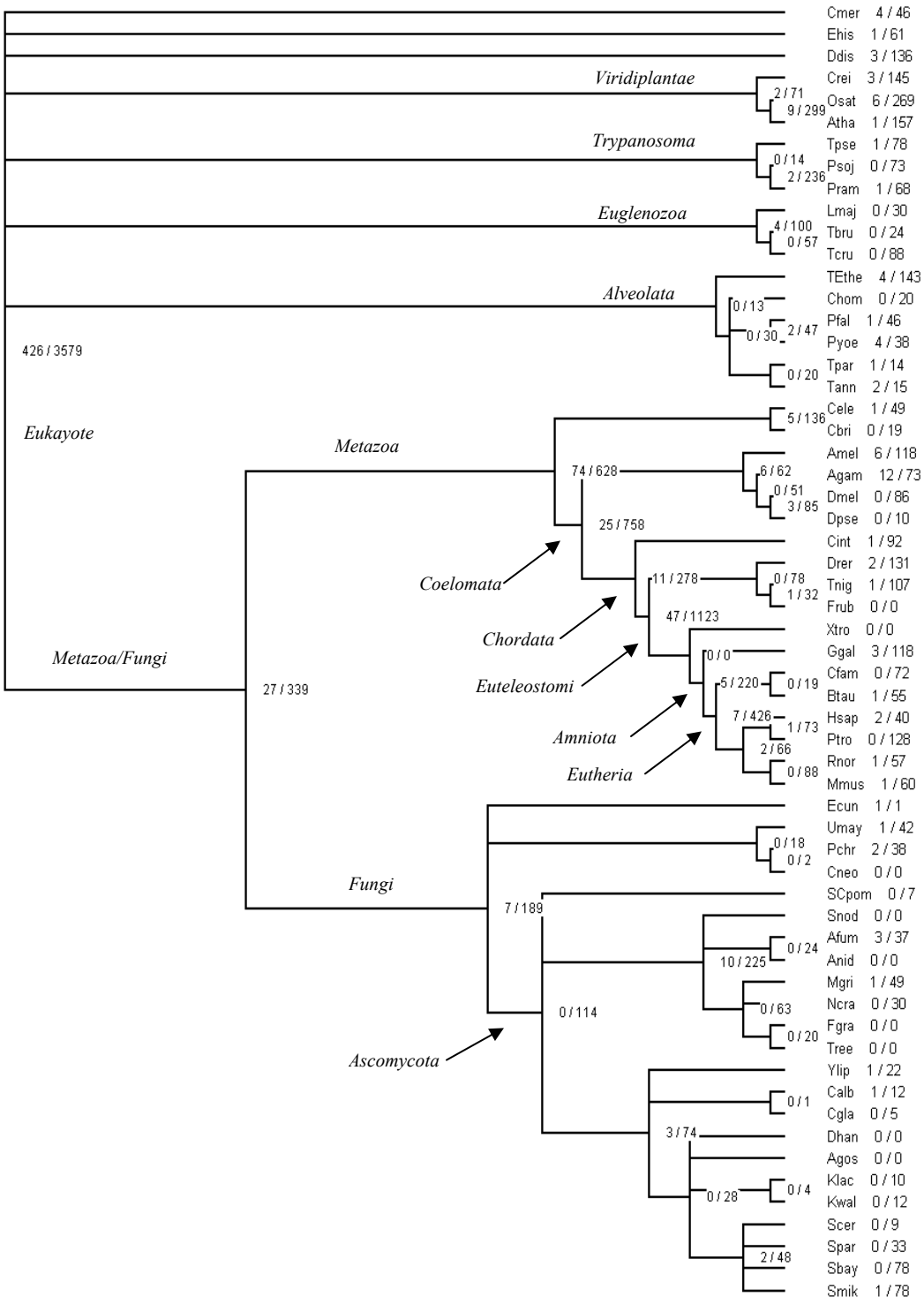


Figure 5.15 The predicted number of domains and domain combinations that originated at each node on the eukaryotic tree.

5.3.3 Conclusion

In summary, the evolutionary history of protein domains can be derived based on their distributions in existing organisms. The evolution of protein domain repertoire can be used to study the evolution of species in general.

5.4 References

1. Koonin, E.V., Y.I. Wolf, and G.P. Karev, *The structure of the protein universe and genome evolution.* Nature, 2002. **420**(6912): p. 218-23.

2. Doolittle, R.F., *The multiplicity of domains in proteins.* Annu Rev Biochem, 1995. **64**: p. 287-314.

3. Ponting, C.P. and R.R. Russell, *The natural history of protein domains.* Annu Rev Biophys Biomol Struct, 2002. **31**: p. 45-71.

4. Rost, B., *Did evolution leap to create the protein universe?* Curr Opin Struct Biol, 2002. **12**(3): p. 409-16.

5. Copley, R.R., et al., *Protein domain analysis in the era of complete genomes.* FEBS Lett, 2002. **513**(1): p. 129-34.

6. Chothia, C., et al., *Evolution of the protein repertoire.* Science, 2003. **300**(5626): p. 1701-3.

7. Vogel, C., et al., *Structure, function and evolution of multidomain proteins.* Curr Opin Struct Biol, 2004. **14**(2): p. 208-16.

8. Teichmann, S.A., J. Park, and C. Chothia, *Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14658-63.

9. Apic, G., J. Gough, and S.A. Teichmann, *Domain combinations in archaeal, eubacterial and eukaryotic proteomes.* J Mol Biol, 2001. **310**(2): p. 311-25.

10. Vogel, C., et al., *Supra-domains: evolutionary units larger than single protein domains.* J Mol Biol, 2004. **336**(3): p. 809-23.

11. Vogel, C., S.A. Teichmann, and J. Pereira-Leal, *The relationship between domain duplication and recombination.* J Mol Biol, 2005. **346**(1): p. 355-65.

12. Kummerfeld, S.K. and S.A. Teichmann, *Relative rates of gene fusion and fission in multi-domain proteins.* Trends Genet, 2005. **21**(1): p. 25-30.

13. Gough, J., *Convergent evolution of domain architectures (is rare).* Bioinformatics, 2005. **21**(8): p. 1464-71.

14. Vogel, C. and C. Chothia, *Protein family expansions and biological complexity.* PLoS Comput Biol, 2006. **2**(5): p. e48.

15. Orengo, C.A., et al., *CATH--a hierarchic classification of protein domain structures.* Structure, 1997. **5**(8): p. 1093-108.

16. Orengo, C.A., et al., *The CATH protein family database: a resource for structural and functional annotation of genomes.* Proteomics, 2002. **2**(1): p. 11-21.

17. Buchan, D.W., et al., *Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database.* Genome Res, 2002. **12**(3): p. 503-14.

18. Ranea, J.A., et al., *Evolution of protein superfamilies and bacterial genome size.* J Mol Biol, 2004. **336**(4): p. 871-87.

19. Ekman, D., et al., *Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions.* J Mol Biol, 2005. **348**(1): p. 231-43.

20. Bjorklund, A.K., et al., *Domain rearrangements in protein evolution.* J Mol Biol, 2005. **353**(4): p. 911-23.

21. Bjorklund, A.K., D. Ekman, and A. Elofsson, *Expansion of protein domain repeats.* PLoS Comput Biol, 2006. **2**(8): p. e114.

22. Trifonov, E.N. and I.N. Berezovsky, *Evolutionary aspects of protein structure and folding.* Curr Opin Struct Biol, 2003. **13**(1): p. 110-4.

23. Fitzkee, N.C., et al., *Are proteins made from a limited parts list?* Trends Biochem Sci, 2005. **30**(2): p. 73-80.

24. Deeds, E.J., N.V. Dokholyan, and E.I. Shakhnovich, *Protein evolution within a structural space.* Biophys J, 2003. **85**(5): p. 2962-72.

25. Shakhnovich, B.E., et al., *Protein structure and evolutionary history determine sequence space topology.* Genome Res, 2005. **15**(3): p. 385-92.

26. Hou, J., et al., *Global mapping of the protein structure space and application in structure-based inference of protein function.* Proc Natl Acad Sci U S A, 2005. **102**(10): p. 3651-6.

27. Mader, S., *Biology*. eighth ed. 2004, New York: McGraw-Hill.

28. Veretnik, S., et al., *Toward consistent assignment of structural domains in proteins.* J Mol Biol, 2004. **339**(3): p. 647-78.

29. Ochman, H., J.G. Lawrence, and E.A. Groisman, *Lateral gene transfer and the nature of bacterial innovation.* Nature, 2000. **405**(6784): p. 299-304.

30. Brown, J.R., *Ancient horizontal gene transfer.* Nat Rev Genet, 2003. **4**(2): p. 121-32.

31. Olsen, G.J., *The history of life.* Nat Genet, 2001. **28**(3): p. 197-8.

32. Doolittle, W.F., et al., *How big is the iceberg of which organellar genes in nuclear genomes are but the tip?* Philos Trans R Soc Lond B Biol Sci, 2003. **358**(1429): p. 39-57; discussion 57-8.

33. Welch, R.A., et al., *Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli.* Proc Natl Acad Sci U S A, 2002. **99**(26): p. 17020-4.

34. Zhaxybayeva, O., et al., *Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events.* Genome Res, 2006. **16**(9): p. 1099-108.

35. Bapteste, E., et al., *Phylogenetic reconstruction and lateral gene transfer.* Trends Microbiol, 2004. **12**(9): p. 406-11.