ABSTRACT

Title of Document: **INVESTIGATION OF SOME POSSIBLE ORIGINS OF PROTEIN FAMILIES**

Nuttinee Teerakulkittipong, Ph.D., 2013

Directed By: **Professor John Moult,**
Institute for Bioscience and Biotechnology Research
Department of Cell Biology and Molecular Genetics

The prevailing view of the evolutionary history of proteins has been that all protein domains are descendents of distinct evolutionary lines, and that these lines are all relatively ancient families. The primary basis for that view was that known protein structures could be grouped by similarity of topology into a small number of folds. However, two lines of evidence challenge that view of protein evolution. First, analysis of sequence relationships within and between sets of complete genomes has established that a large proportion of protein sequence families are narrowly distributed in phylogenetic space and so appear to be relatively recent in origin. Second, analysis of the relationship between known protein structures shows that there are many more than a 1000 distinct folds, appearing to imply many more evolutionary lines. There are four hypotheses for the discrepancy between the traditional view and the observed structural and sequence distributions within protein families. Specifically, these are that apparently young protein families may arise from (1) previously non-coding DNA, or frame-shifted from existing coding sequence, (2)

recombination of structural fragments between proteins or recombination with non-coding DNA, (3) older families where the rapid rate of sequence change makes relatives hard to detect, and (4) lateral gene transfer (LGT) from other organisms. In the investigation of these hypotheses, phylogenetic analysis provides a means of estimating the relative age of protein families and of detecting lateral gene transfer effects. Phylogeny based investigation of prokaryotic species divergence has generally been performed using a small number of families resulting in significant bias that affects age analysis. Therefore, we decided to use information from many protein families for constructing a species tree, utilizing a new procedure for combining these diverse sources. The resulting tree for 66 Prokaryotic species incorporates information from 1,379 protein families. The families were selected on the basis of consistent family evolutionary rates obtained using three different methods. Noise resistant methods were used to combat the effects of lateral gene transfer and some inevitable errors in protein sequence alignment and identification of orthologous families. Most topological features of the tree are robust as assessed by bootstrap testing, and previous distortions of inter-kingdom distances and poor determination of short branch lengths have been corrected. The tree is used to obtain estimates of the age of all protein families, key to the investigation of all four hypotheses. Proteins affected by LGT events were detected using a previously developed method, and removed before the age calculation.

We used the estimated family ages obtained from the phylogenetic analysis to examine five properties of proteins as a function of the age of the corresponding

families. The goal here is to ascertain whether the age dependence of these properties supports hypotheses (1) and (2) for the origin of apparently young families – that is, these are truly new open reading frames. The five properties are the mRNA expression level, relative evolutionary rate, predicted percentage of structural disorder, number of protein interaction partners and codon composition bias. The results are consistent with the new open reading frame model: Expression is found to increase substantially as a function of family age, suggesting that young proteins are not yet adapted sufficiently to tolerate high concentration conditions. The rate of change of amino acid change is faster for young proteins, consistent with overall positive selection for improved structural and functional properties. The fraction of predicted disorder is highest in the youngest proteins, consistent with immature structural properties. The number of known protein-protein interactions increases steadily with age, with low levels for young proteins, suggesting an ongoing process of increasing functional complexity. Analysis of these four factors is reported in Chapter 3.

Results for the final factor, codon compositional bias, are reported in Chapter 4. Here we found that the codon composition of young proteins is markedly different from that of old proteins and similar to that of proteins constructed with random codon assignment. Thus the results are consistent with a model of many young proteins having newly formed open reading frames, and that during the subsequent evolution process, the codon composition is gradually optimized to fit the specific genomic conditions of the organism concerned.

Overall, results for all five properties lend statistical support to the new open reading frame hypotheses. Further investigation is needed however. In particular, examination of the structural properties of young proteins, such as super-secondary structure composition and the distribution of use of rare and common structural fragments, should be useful.

INVESTIGATION OF SOME POSSIBLE ORIGINS OF PROTEIN FAMILIES.

By

Nuttinee Teerakulkittipong

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:
Professor John Moult, Chair
Professor Catherine Fenselau
Professor James Culver
Associate Professor Kevin Mclver
Professor Leslie Pick

Dedication

To my family

Acknowledgements

My foremost gratitude is dedicated to my academic advisor, Professor John Moult. I still remember six years and a half ago when I started my third lab rotation in John's group, I was very new to computational biology. He led me into this fascinating world of protein evolution and made enthusiastic me about my project. His guidance has made this a thoughtful and rewarding journey. I appreciate very much his great insight and guidance for my Ph.D. work, his great sense of humor, his energetic and nice way of demanding high quality work. Without him, I can't imagine how I achieved so much after six years' study.

I also would like to express my appreciation to my candidacy and dissertation committee members: Professor Catherine Fenselau, Professor Leslie Pick, Associate Professor Kevin Mclver and Professor James Culver. They are knowledgeable, professional and kind. They have given me a lot of excellent advice. This work won't be the same without their help and encouragement.

I would like to acknowledge all the help from Dr.Lipika Ray, a Post-doc in the Moult group. She provided needed encouragement and insight, especially in her critical reading of my paper and thesis. I would like to thank to all people in the Moult group: Dr. Eugene Melamud, Xijun (Ethan) Zhang, Dr. Zhen Shi, Maya Zuhl, Albert (Chen-Hsin) Yu, Chen Cao, and Yizhou Yin, for much valuable scientific discussion and for sharing their personal lives with me.

I also own great thank to my parents for their endless love, support and sacrifice. Although far from here, they have given me priceless support. I know their hearts are

always with me. My special thanks is for my husband, Mr. Worawit Teerakulkittipong. I thank him for always being there to support me.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Section 1 An Integrated View of Protein Evolution, the Presence of Single-Member Families and the Possible Origin of Young Proteins

The prevailing view of the evolutionary history of proteins has been that all belong to a relatively small number of ancient families. Chothia[1] argued that there are about 1000 such families. The latest version of the Structural Classification of Proteins (SCOP) 1.75 release (June 2009)[2], containing 110,800 domains and with structures organized into 3,902 families, 1,962 super-families and 1,195 folds, also supports this. The rapid accumulation of new structural data and rapidly increasing knowledge of the complete genome sequences provides a basis for a broader based analysis. As a result, two lines of evidence, one based on structure and the other on sequence, now suggest the traditional view of protein evolution is not correct.

Subsection 1. Views of Protein Emergence and Change

There have been several more recent analyses of the accumulation of structural diversity in the Protein Data Bank (PDB)[3, 4] usually suggesting that there are many more than a 1000 natural folds that fit the SCOP definition. Previous work in our lab[5] classified folds into three classes: superfolds, which are adopted by very many protein families and are highly recurrent within proteomes; mesofolds, which have an intermediate number of protein families associated with them; and unifolds, found for

single narrowly distributed sequence families. The resulting estimate is that there are at least 10,000 folds, and probably many more (figure 1).

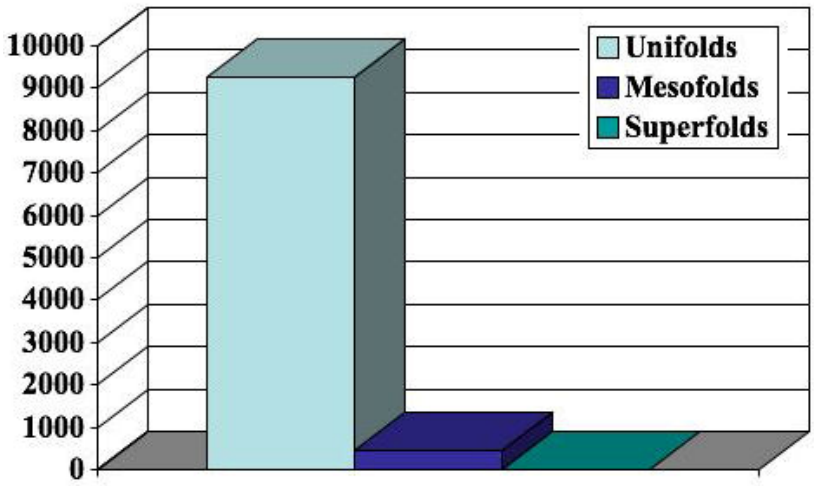


Figure 1. Distribution of protein fold use in biology[5]*. There are a large number of folds narrowly distributed in sequence space (unifolds, left bar, few with structure), a moderate number of folds found in a few sequence families (mesofolds, center bar, most with structure), and a very small number of very common folds (superfolds, right bar, all with structure).*

Large-scale genome sequencing projects enable us to analyze the sequence relationships within and between sets of complete genomes. One interesting finding is that a substantial percentage of each newly sequenced genome consists of protein coding ORFs (Open reading frames) that do not resemble any other sequences in the sequence databases. Some of these families have still so far been found in only a single genome, and have only one member, and so are often referred to as singletons or ORFans[6, 7]. An analysis of 66 bacterial and archaeal genomes[8] found that 20,992

of the protein families apparently have members in only a single one of these genomes, about two-thirds of the total (figure 2). Thus, as with structure, the sequence view shows most protein families are narrowly distributed in phylogenetic space, and so apparently of recent origin, suggesting the continuous emergence of new, independent evolutionary lines.

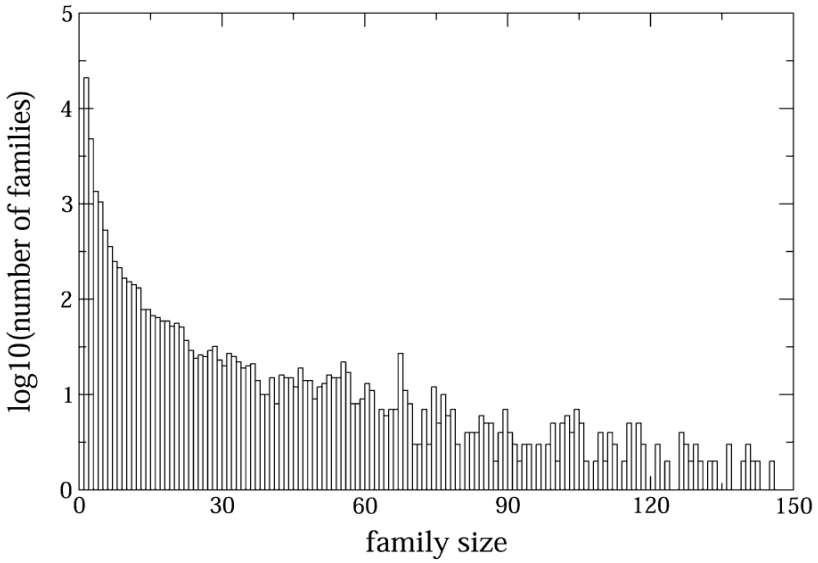


Figure 2. Distribution of domain family size in a set of 66 genomes[8]. There is an approximately power-law relationship between family size and the number of families, with very many small families, and only a few large ones. There are 20,992 singletons (families with only one member), about 2/3 of the total, and 4,810 doubletons (family size 2). At the other end of the spectrum, there are 263 families larger than 100.

Apparent singleton genes and proteins have been investigated across other sets of organisms, revealing different fractions of these in different species and phyla, as

illustrated in the figure 3. Across primates (human, chimpanzee and macaque), around 1.3 % of genes are singletons[9]. In Fungi, Drosophila, Nucleocytoplasmic large DNA viruses, Rickettsia genus and another microbial genome the fraction of singletons is 2.0%, 18%, 2.8%, 30.0% and 14% respectively[10, 11, 12, 13, 14]

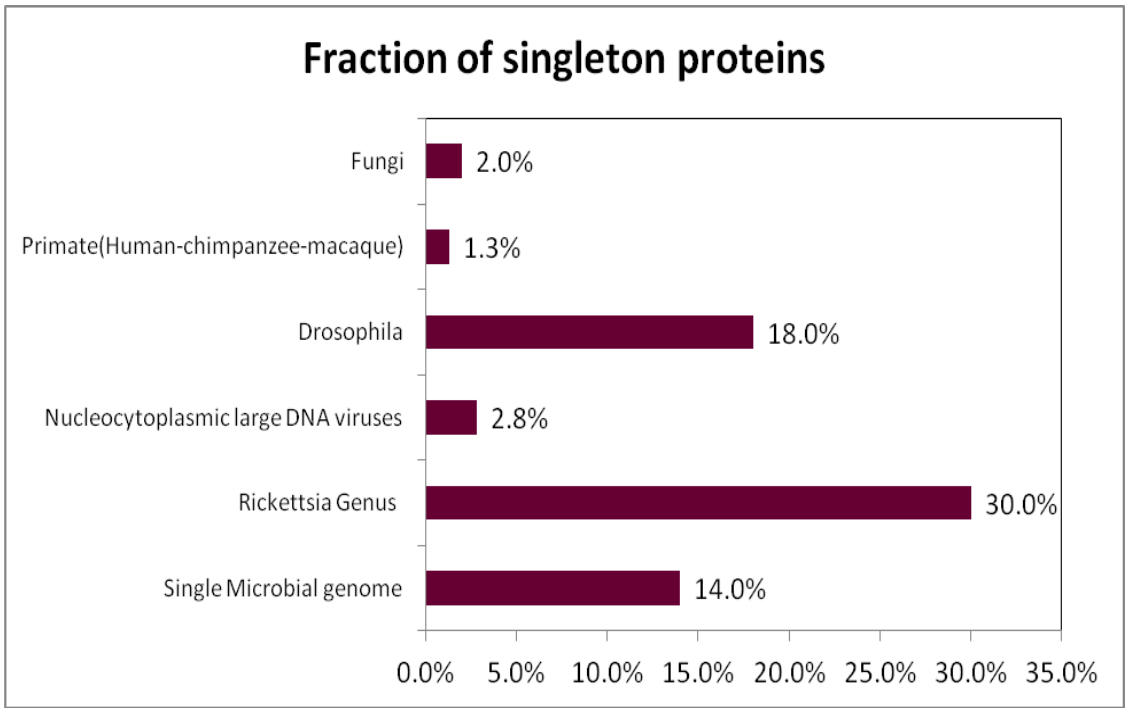


Figure 3. Singleton Proteins in other organisms and orders.

The presence of so many singletons suggests that protein diversity in nature may be greater than previously expected. However, because little can be learned about singletons via homology, each of them represents a mystery, awaiting interpretation. All of these new data provide us the opportunity to examine possible explanations for the discrepancy between prevailing wisdom and experimental fact, and so produce a new and coherent model of protein evolution.

Subsection 2. Possible Explanations of the origin of apparently young proteins

The mechanisms that lead to the emergence of young proteins are not fully understood. Several possible explanations were given in many previous studies for this phenomenon, including by us. As part of an NIH funded project (R01GM81511, Mechanisms of Protein Structure Evolution), we have proposed four possible hypotheses to account for the wealth of phylogenetically narrowly distributed proteins:

1. Apparently young proteins are coded for by new genes, formed from previously non-coding DNA, or frame-shifted from existing coding sequence. If it is true that new open reading frames play a significant role in generating young protein families, it should be possible to identify cases where this has occurred. Some examples from Eukaryotes are known, for instance a recently evolved antifreeze protein, originating from a short partly intronic sequence[15]. A more general study has found over a 1,000 instances of intronic sequences converting to exons in the period between the divergence of Human and rodent[16]. Additionally, a study in Drosophila has identified five new *D. melanogaster* genes that are derived from noncoding DNA[17]. These limited examples provide indirect support for the explanation that many apparently young proteins in prokaryotes emerged from non-coding regions of each genome.

2. Protein structure changes continuously, through a process of local conformational change, recombining structural fragments between proteins, and recombination with non-coding DNA, so that distant evolutionary relationships are unrecognizable at the structure level. The composite nature of proteins is well established, with components

ranging from small indels of a few residues to mixing and matching of semi-autonomous domains. Shuffling of complete domains has been extensively analyzed[18], and proposed as a primary mechanism for the emergence of new function, particularly in eukaryotes.

3. Apparently young proteins are a result of lateral gene transfer from other organisms. Lateral gene transfer (LGT), also called horizontal gene transfer, is the process of transfer of genes between different species. There are several LGT mechanisms, such as transduction by viral and phage genomes and conjugation with exchanging their plasmid[22].

4. Apparently young protein families are in fact often much older, but rapidly evolving rates of sequence changes make relatives hard to detect. This possibility has been suggested by Long et al[21].

In this work, we have investigated the first hypothesis, that apparently young proteins are coded for by new genes, formed from previously non-coding DNA, or frame-shifted from existing coding sequence. To this end, we have examined five relevant protein properties: protein expression level, relative evolutionary rate, number of protein-protein interaction partners, predicted intrinsic disorder region, and codon usage; as a function of age. Chapters 3 and 4 describe this work.

Section 2 Studies of Prokaryotic Species Trees

Subsection 1. Reconstruction of Phylogenetic trees of Prokaryotic Organisms

Phylogenetic analysis of DNA or protein sequences has become an important tool for studying the evolutionary history of organisms from bacteria to humans. Since the rate of sequence evolution varies extensively over genes and DNA segments[23, 24], one can study the evolutionary relationships of virtually all levels of classification of organisms by using different genes or proteins. There are many statistical methods that can be used for reconstructing phylogenetic trees from molecular data[25]. The true tree is almost always unknown, and it is difficult to test the accuracy of the trees obtained by different tree building methods. Temporal information concerning prokaryote evolution has come from diverse sources and is difficult to integrate due to a limited fossil record and the complexities associated with the molecular clock and deep divergences. For instance, phylogenetic analysis of genes, and, more recently, information contained in completely sequenced genomes, contribute to our view of how widespread LGT must be in evolution. Interpretations of these data have led to arguments that rampant LGT would erase phylogenetic history especially in terms of changing protein family age[26, 27, 28]. Early work focused on building trees using sequence relationships between orthologous ribosomal 16s RNA genes, which are ancient and distributed over all lineages of life with little or no lateral gene transfer[29] , for example resulting in a ribosomal RNA based tree covering the three domains of life including the two prokaryotic kingdoms[29]. Therefore rRNAs are commonly recommended as the principal molecular phylogenetic marker[26]. However, the opposing view is that 16s rRNA genes can lead to erroneous tree topology as unrelated phylogenetic relationships are placed close in phylogenetic trees due to similarity in nucleotide composition of evolutionarily distant 16s rRNA genes[30]. As a

result, many researchers turned to protein coding genes, such as in the study of metagenomic bacterial ecology[30]. Phylogenetic analyses of protein amino acid sequences are in general less prone to the nucleotide compositional bias seen in 16s rRNA gene and in protein coding genes[30, 31]. The evolutionary history of prokaryotic species divergence has previously been investigated using protein families that have members in all fully sequenced genomes (21 - 31 protein families)[26, 32, 33, 34, 35, 36, 37].

In our lab, Yongpan Yan built a reference tree with distances derived from the average sequence identities over a set of fourteen conserved orthologous protein families (most are ribosomal proteins) that have members in each of 66 prokaryotic genomes. He used this reference tree to obtain a preliminary estimate of the extent of LGT, and found that 18% of the genes have undergone transfer within their orthologous family[38]. Analysis of this tree shows some deficiencies. In particular, intergenome distances between some strains of bacteria are related by very short branch lengths and intergenome distances between bacteria and archaeal species are too long[38]. This result suggests that ribosomal proteins are atypical in a number of respects.

Determination of the relationship between species using phylogenetic trees based on a single or small set of genes or proteins encounters three main problems: a limited number of sequences, variability of evolutionary rates in different lineages, and the effect of lateral gene transfer. The first two factors add uncertainty to tree reconstruction; the last factor leads to protein phylogenies being genuinely different from species phylogeny. The determination of complete genome sequences of many bacteria and archaea created the opportunity for a new level of phylogenetic analysis

that is based not on a phylogenetic tree for selected molecules but rather on the entire body of information contained in the genomes or on a rationally selected, substantial part of this information[32]. We expect the topology and branch lengths of such trees to be better determined than for trees based on few families. Since these properties are important in studying the protein family age, we decided to construct a prokaryotic species tree, utilizing information from all protein families. Chapter 2 describes this work.

Subsection 2. Lateral gene transfer

Lateral gene transfer, also called horizontal gene transfer, is a process whereby genetic material contained in small packets of DNA can be transferred between individual organisms[38]. For many years it was the common belief that lateral gene transfer was rare, and did not play a significant role in evolution. As sequence-based genomics has developed, it has become more and more obvious that the process is very common and plays an important role in evolution[39]. There are three possible mechanisms of LGT (figure 4). These are transduction, transformation and conjugation. Transduction occurs when bacteria-specific viruses or bacteriophages transfer DNA between two closely related bacteria. Phages can exchange genes with their hosts, by integrating them as prophages or by exchanging individual genes with their hosts via recombination[40, 41]. Phages exchange genes with other phages mostly when they are inside the same host cell and with prophages residing in the host genome[42, 43].

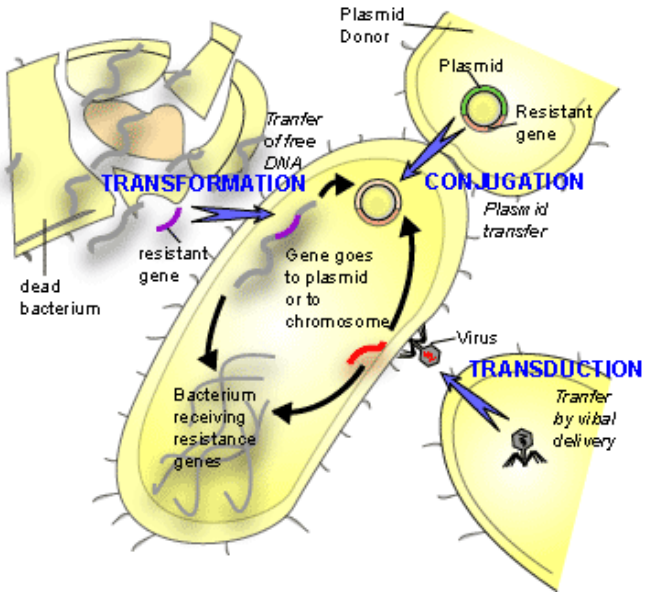


Figure 4. Lateral gene transfer possible mechanisms taken from[44] *(Yim, G. 2009).*

Transformation is a process where parts of DNA are taken up by the bacteria from the external environment. This DNA is normally present in the external environment due to the death of another bacterium. Conjugation occurs when there is direct cell-cell contact between two bacteria (which need not be closely related) and transfer of small pieces of DNA called mobile genetic elements takes place, including plasmids, transposons, integrons, and other integrative conjugative elements (ICEs) that mediate the movement of DNA within genomes and between genomes[45, 46, 47]. It has been estimated that between 8% and 18% of the *E.coli* genome was acquired by lateral gene transfer[41].

In other genomes, the estimated extent of transfer varies over a wide range, from almost none in small genomes such as *Mycoplasma genitalium*, *Rickettsia prowazekii*

and *Borrelia burgdorferi*, to about 24% in *Thermotoga maritime*[48, 49]. Studies have shown that lateral gene transfer events can happen across large phylogenetic distances, for example, isoleucyl-tRNA synthetases, whose acquisition from eukaryotes by several bacteria is linked to antibiotic resistance[50]. Clearly, the mechanisms for transferring genes in nature are abundant, but the frequency with which these elements overcome barriers to transfer to attain successful integration into new environments needs further elucidation and is still under debate[46]. In this project, we are interested in LGT because these events may cause protein family ages to appear larger than they really are. For this reason, in Chapter 3, we identify likely LGT events within the prokaryotic kingdom and eliminate the transferred genes from our calculation of family ages.

Section 3 Properties of proteins as a function of age

The hypothesis that most young proteins are composed of newly created open reading frames implies a number of properties may be different between young and old proteins. We explored five of these properties: the level of mRNA expression, the relative rate of amino acid sequence change within a family, the level of structural disorder, the number of protein-protein interactions, and the codon usage.

Subsection 1. The estimation of relative age of each orthologous protein family

All genomes are collections of genes and proteins that widely differ with respect to their histories and their intrinsic characteristics. In many studies the age of a protein is

mostly defined by considering the taxonomic distribution of the proteins in the family, analyzing the presence or absence of members in diverse lineages[51, 52, 53, 54, 55, 56, 57]. For example, some proteins in an organism are ‘‘old,’’ in the sense that they have identifiable orthologs across a diverse range of species spanning vast evolutionary distance. Other proteins are ‘‘young’’ in the sense that orthologs are identifiable only in one species or closely related species[51, 52, 53, 54, 55, 56, 57]. For our study, we deduce the age of each orthologous protein family from the species tree constructed in Chapter 2), using the method described in Chapter 3.

Subsection 2. mRNA Expression level as a function of family age

A previous study in yeast[56] observed a positive correlation between mRNA expression level and gene age. We also observed sharp increase in expression level from the youngest to the oldest families, as described in Chapter 3. It was also observed previously that there is a strong negative correlation between gene expression level and the rate of sequence change in some organisms, such as yeast (*Saccharomyces cerevisiae*)[55, 56, 58] and E.coli[59]. Drummond and Wilke have proposed this correlation is a consequence of mutations in more highly expressed genes having a greater effect on fitness, as a result of being more prone to causing aggregation or overwhelming the chaperone machinery[61, 62, 63]. This phenomenon has also been explained by Yang et al. in term of mutations in highly expressed proteins being more likely to result in incorrect interactions that are wasteful and potentially toxic[64]. As described in Chapter 3, we observed that the apparent correlation between E.coli K12

mRNA expression and family evolutionary rate is in fact an artifact of both these quantities correlating with age.

Subsection 3. Protein family evolutionary rate and relative family age

Variation in the rate and pattern of amino acid substitution in proteins is a fundamental property of protein evolution. The rate of amino acid substitution varies considerably among different protein families[65]. Changes in protein sequences are constrained by selection pressure and so accumulate at different rates[53]. For higher Eukaryotes such as human, fugu, fly and worm, it has already been reported that young proteins are under strong positive selection[53, 56, 65]. Young proteins evolve under variable selection pressure and their evolutionary rates are faster than older proteins[53, 54, 55, 66]. We find a strong decreasing trend for evolutionary rate as a function of increasing age for E. coli K12 proteins, as described in Chapter 3.

Subsection 4. Correlation of number of protein- protein interactions and family age

In recent years, with the explosive development of high-throughput experimental technologies, the number of reported protein–protein interactions (PPIs) has increased substantially. Large collections of PPIs produce "omic" scale views of protein partners and protein membership in complexes and assemblies in many organisms[67] . However, there are very few publications that investigate the relationship between physical protein-protein interactions and age of the protein. Two studies have been performed in yeast and a strong and positive correlation of protein-protein interaction

and age of a protein was found[68]. Other studies by Kunin et al. also investigated the relationship that proteins of different ages have different connectivity levels in interaction networks[69]. In this work we also observed a steady increase in the number of reported protein-protein interactions for E. coli K12 proteins with increase in family age, as described in Chapter 3.

Subsection 5. Relationship of predicted percentage protein disorder and family age

Many proteins contain regions without well-defined structure (intrinsically disordered regions) and it has been suggested that these are associated with particular functions, including cell regulation, nuclear localization, chaperone activity, antibody creation, signaling, as well as binding to proteins, DNA, and other ligands[70, 71, 72, 73]. Protein disorder is more prevalent in complex organisms, by some estimates accounting for 33 % of the residues in human proteome, but only a few percent of residues in E.coli, leading to the suggestion that it may play a major role in the evolution of complexity[70]. We observed a steady decrease in predicted structural disorder for E. coli K12 proteins with increasing age, followed by slight increase again for the oldest subset of families, as described in Chapter 3. It has been observed that disorder increases with the number of protein interactions[71, 72, 73, 74], and we suggest the cause of the late age increase is that as the number of interactions increases, segments of proteins become more disordered to allow interaction with multiple partners.

All of these protein properties correlate well with family age, but mere correlation does not establish a causal relationship with age. In order to better understand which

underlying effects cause which observations, we performed a set of partial correlation analyses[75], examining the effect of removing the influence of each factor on correlations between each pair of variables for E.coli K12 proteins, as explained in Chapter 3.

Subsection 6. Composition bias in different organisms and its relationship to protein age

Codon usage bias refers to differences in the relative frequency of occurrence of synonymous codons in coding DNA. The redundancy in the number of codons for most amino acids can result in different codon compositions in different organisms[77]. How these organism specific preferences arise is a much debated area of molecular evolution[76, 77, 78]. Different factors have been proposed as related to codon usage bias, including gene expression level (reflecting selection for optimizing the translation process with respect to tRNA abundance), %G+C composition (reflecting horizontal gene transfer or mutational bias), amino acid conservation, transcriptional selection, RNA stability, optimal growth temperature and hypersaline adaptation[78, 79, 80]. To investigate further the hypothesis that ORFan proteins originate from non-coding regions, we explored the codon composition bias of 47 prokaryotic organisms by comparing the composition bias of the set of all proteins, of random proteins, and of ORFan proteins in each genome. We investigated the codon usage of ORFan proteins with the evolutionary age of the ORFans, and we discuss the evolutionary implications of this observation in Chapter 4.

Chapter 2: Construction of Phylogenetic trees using complete genome information

Section 1 Abstract

Knowledge of complete genome sequences for many organisms provides an opportunity to assess phylogenetic relationships between species on a much broader basis than previously possible. In particular, combining information from the phylogenetic history of many genes may yield a less biased view of species phylogeny. Appropriate combinations of genes can also be used to study the evolution of particular processes. Utilizing these pan-genome data requires the development of new methods that effectively combine information from sequence relationships across a large number of protein families. In turn, combining these data requires estimates of the relative rate of sequence change among families. Particularly among prokaryotes, ambiguities from possible lateral gene transfer events, as well as issues with correctly identifying orthologous relationships and potential errors in sequence alignments necessitate the use of noise resistant methods.

Three noise resistant methods have been used to estimate the relative evolutionary rates of amino acid change within orthologous protein families: least median squares, a Gaussian kernel estimator, and an iterative outlier filtering procedure. Families where the three methods gave consistent rates were normalized to a common rate scale. Intergenome distances were then estimated using the average amino acid substitutions per site information for these families together with the three noise

resistant methods. Standard neighbor joining methods were then used to build a phylogenetic tree from these distances.

Relative evolutionary rates were determined for 2,262 orthologous families extracted from a set of 66 prokaryotic genomes. Rates span a range of about two orders of magnitude, with highest rates typically found for small, phylogenetically narrow families. Data for the 1,379 orthologous families with consistent rates determined by the three different methods were used to estimate the set of all intergenome distances, and these distances in turn were used to obtain a species tree. Bootstrap testing with a 1000 replicates found 75% of the nodes to be determined with 95% or better confidence, and only 10% to be below 50% confidence. Comparison of the tree topology with that obtained using information from a small number of protein families shows a high level of overall agreement, but with specific differences, including separation of the three included bacterial hyperthermophiles in the new tree. Relative branch lengths are also different, particularly showing reduced separation of the bacterial and archaeal kingdoms, as a consequence of reduced reliance on ribosomal proteins. Overall, the results demonstrate the potential of including many protein families in phylogenetic analysis, and in future, choosing sets of families appropriate to a particular biological question.

Section 2 Introduction

Phylogenetic analysis of the relationships between species using molecular and genome level data has become an important tool for studying the evolutionary history of organisms from bacteria to humans[25]. In previous work, we built a prokaryote reference tree with distances derived from the average sequence identities over a set of 14 conserved orthologous protein families that have members in each of 66 prokaryotic genomes. In common with other analyses that use protein families with members in all included species, most of these are ribosomal proteins[27, 32, 33, 34, 35, 36]. Analysis of this tree shows two primary deficiencies[37]. First, intergenome distances between some strains in bacteria are related by very short branch lengths. Likely this is a consequence of the rate of sequence change in conserved families being too slow to properly estimate such short distances. Second, intergenome distances between bacteria and archaeal species appear systematically too long compared with the intra-kingdom distances. That likely arises from the extensive differences between bacterial and archaeal ribosomes[81, 82], resulting in correspondingly abnormally large sequence differences between their proteins across the two kingdoms. The determination of complete genome sequences of many bacterial and archaeal species has created the opportunity for a new level of phylogenetic analysis that is based not on a phylogenetic tree for selected molecules but rather on the entire body of information contained in the genomes or on a rationally selected, substantial part of this information. Here we explore a strategy for utilizing these new data.

All gene family based species tree construction methods must contend with difficulties of identifying orthologous families, and of producing reliable multiple sequence alignments. Prokaryotes present additional problems in the construction of species trees. The fossil record is very limited, providing little useful data against which to validate the results[26, 27]. Lateral gene transfer is very extensive[26, 27, 28], resulting in many genes not representing the evolutionary history of the species they are found in. Indeed, it has been argued that rampant LGT has erased phylogenetic history at the molecular level[26, 27, 28]. We address these difficulties in three ways. First, noise resistant methods are used to combine information from multiple families. Second, results from three different noise resistant methods are compared, so identifying those families where consistent results are obtained. Third, the use of many families allows us to reject any doubtful results, and still have a large and representative set with which to build the final species tree.

Tree building methods fall into two main categories: those that build a tree based on a matrix of distances between entities, and those that build a tree directly from a set of features characterizing each entity. For the construction of species trees, features in the latter method are generally the specific bases or amino acids found at each position in a multiple sequence alignment. A search is made over the space of possible trees, as far as possible finding the tree that satisfies some optimization criterion, such as Maximum parsimony (MP)[83], Maximum Likelihood (ML)[83, 84, 85] and maximum posterior probability[86]. These methods are conceptually appealing, and make use of information from all individual substitutions in the sequences included[86, 87]. On the other hand, the methods are very compute intensive, and so

cannot be scaled to include many sequences. As a result, most species trees built to date are based on information from a small number of gene families[27, 32, 33, 34, 35, 36], and do not utilize the wealth of data provided by complete geneome sequences. Distance-based tree building methods estimate the relative pair-wise distances between each pair of species, and construct a phylogenetic tree from the resultant distance matrix usually by the Neighbor Joining (NJ) method[86, 88]. These methods are not computationally demanding, so there is no limit of number of sequences that can considered. In this work, we develop distance-based tree construction methods that take advantage of the complete genome information available for prokaryotic species.

Section 3 Results

Subsection 1.Comparison of family evolutionary rates from different methods

Family evolutionary rates were calculated as described in Section 5 Methods. Least median squares provided solution for 2,403 out of 4,856 orthologous families included, and the Gaussian kernel estimator, 2,264. Most poorly determined rates are for families with less than five members, and arise from insufficiently distinct points. Figure 5 shows a comparison of family evolutionary rates obtained using three different methods for the 2,264 families where all three methods returned a value. There is good agreement between the Gaussian kernel density estimator and least median squares (x axis and y axis), while the recursive filtering method tends to return higher values for families with low rates (*z axis*). The correlation coefficient between the LMS and GKDE is 0.92, between LMS and RF it is 0.57, and between GKDE and RF, 0.56.

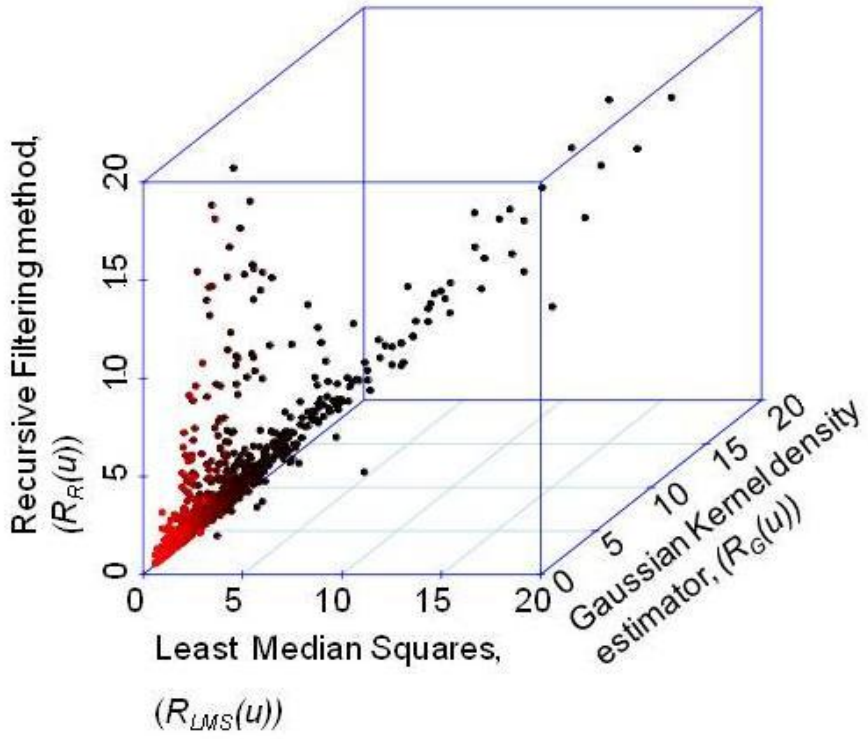


Figure 5. 3D scatter plot showing the comparison of estimated evolutionary rates from three methods: least median squares (x), Gaussian kernel density estimator (y) and recursive filtering with three iterations (z). (Redder points are closer to the x-z plane.)

Of the 2,264 families where the three methods provided values, consistent rates (as defined (in Section 5 Methods) were obtained for 1,379 families. Figure 6 shows the distribution of rates for these families. The peak of the distribution is at slightly higher rate than that of the conserved reference families (relative rate 1.0). There is long tail of significantly higher rates than 4. As figure 7 shows, many of these high rates are for apparently small families with less than 10 members.

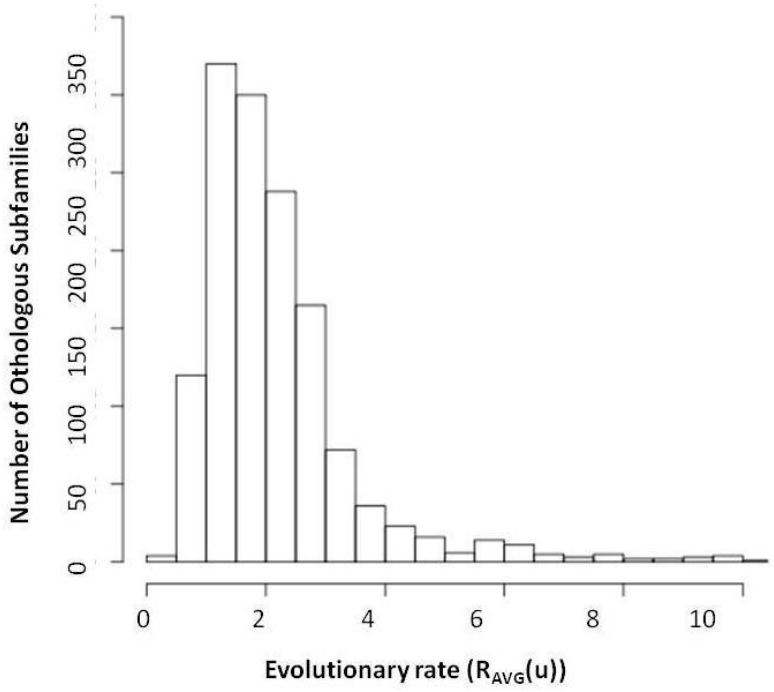


Figure 6. Distribution of protein family relative evolutionary rates of 1,379 families, $R_{AVG}(u)$. A rate of 1.0 corresponds to the average for 14 highly conserved families, with members in all 66 genomes considered.

There are also 124 families with lower rates than the reference ones (see examples in Supplementary figure S1). 33 of these are ribosomal proteins. Among the others with evolutionary rates slower than 1.0, 67 are annotated as metabolism enzymes, or involved in transcription and translation control, sporulation, and cell division (www.ncbi.nlm.nih.gov/protein) and these categories are 3.74 times enriched (chi-square P-value < 0.0001) in the slow rate group compared to the fast rate (> 1.0) group. The remaining 24 slow rate families are annotated as conserved hypothetical proteins or proteins of unknown function. These categories are not significantly enhanced in the slower rate group compared to the fast one (chi-square P-value = 0.12).

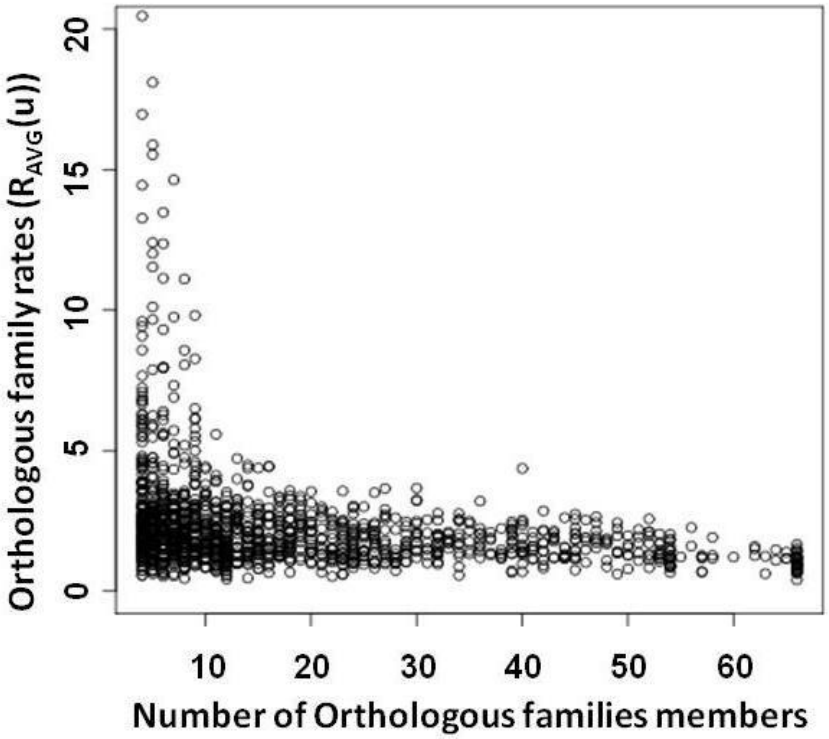


Figure 7. Orthologous family evolutionary rates as a function of family size. Some small families exhibit anomalously high rates.

We also found 85 families where relative rates are greater than 5.0 compared to the 14 conserved protein families, with maximum of 42. As figure 7 shows, all these high rate families have less than 10 members.

Subsection 2. Comparison of intergenome distances derived with different methods

Multi-family based intergenome distances, *D(i,j)* between each pair of species *'i'* and *'j'* were calculated with the three noise resistant methods including all contributions from the 1,379 protein families with consistent evolutionary rates. Comparison of the

intergenome distances derived with the three different methods (figures 8 and 9) shows generally high agreement. More specifically, 96.9% of all intergenome distances satisfy the less than 20% deviation criterion set out in Section 5 Methods. The remaining 67 genome pairs are from very closely related species. For these, intergenome distances as calculated by the LMS method are used in the subsequent tree building.

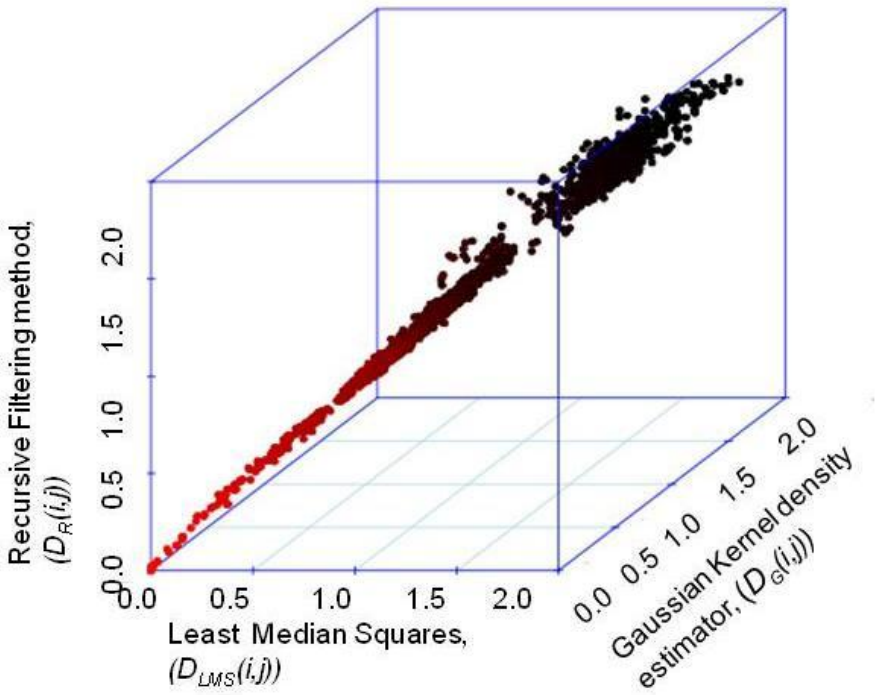
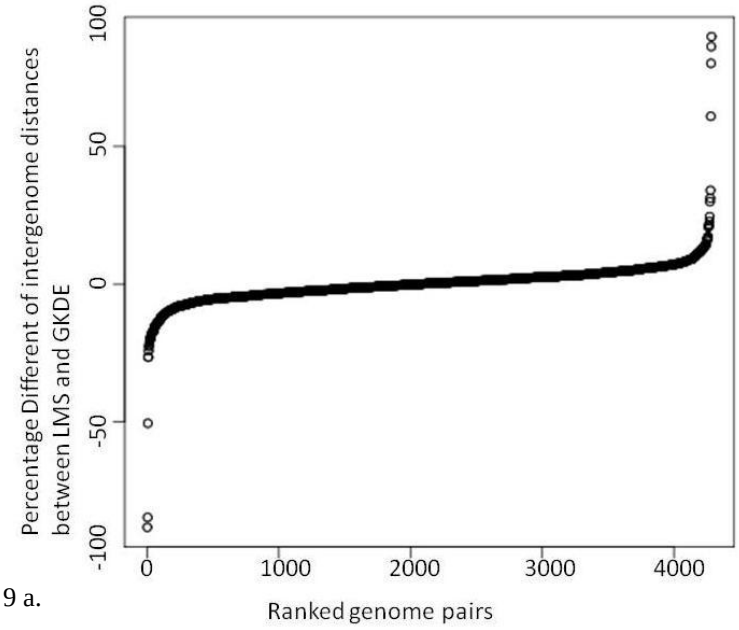
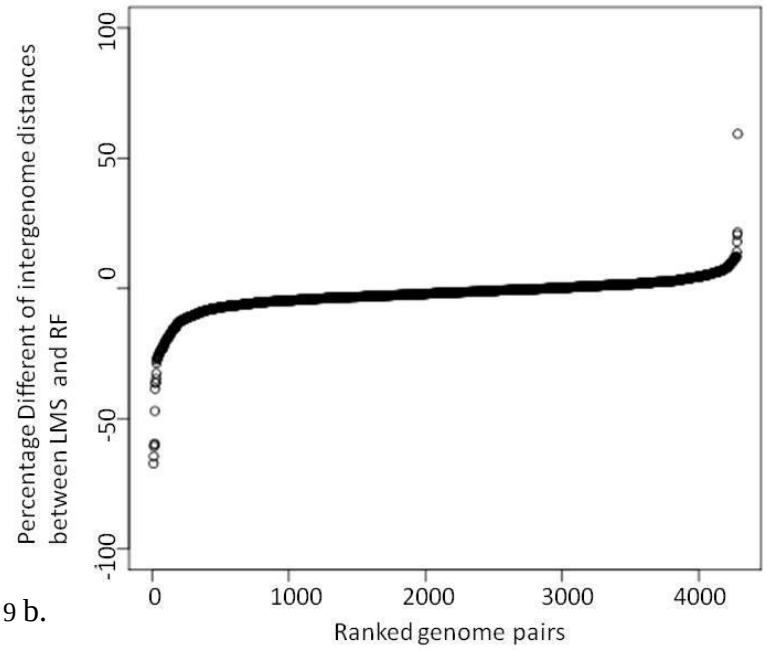


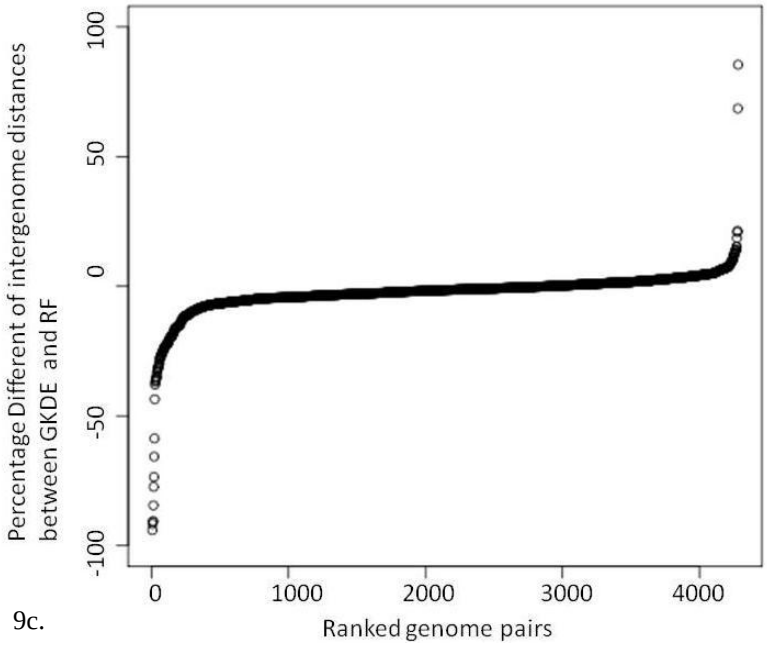
Figure 8. Comparison of inter-genome distances derived using three different methods: least median squares (x), Gaussian kernel density estimator (y) and the recursive filtering (z). (Redder points are closer to the x-z plane.)


Percentage Different of intergenome distances between LMS and GKDE
100
50
0
-50
-100
0
1000
2000
3000
4000
Ranked genome pairs

9 a.


Percentage Different of intergenome distances between LMS and RF
100
50
0
-50
-100
0
1000
2000
3000
4000
Ranked genome pairs

9 b.



9c.

Figure 9. Distribution of percentage different in intergenome distances derived from three methods: least median squares (LMS) and the Gaussian kernel density estimator (GKDE) (9a.), least median squares and recursive filtering (RF) (9b.) and the Gaussian kernel density estimator and recursive filtering methods (9c). Agreement between methods is generally high.

Subsection 3. Comparison of intergenome distances obtained with a few versus many families

Figure 10 shows a comparison of inter-kingdom distances derived with the data from only the set of 14 conserved families and those obtained using information from 1,379 families. The result shows a consistent reduction in inter-kingdom distances in the new set of distances: The average fractional change in inter-kingdom distances

between the new multifamily intergenome distances and the previous 14 family set is -0.081 consistent with reduction of the bias from ribosomal proteins.

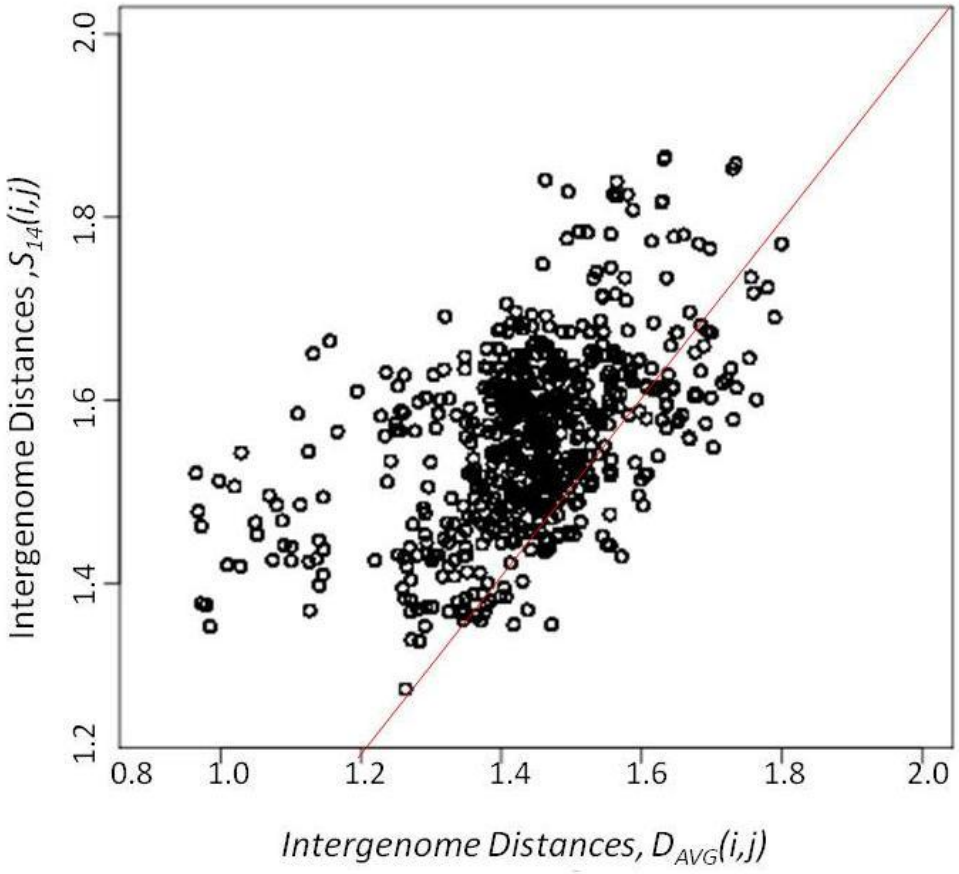


Figure 10. Comparison of inter-kingdom distances obtained using many families (X axis) and only using 14 conserved protein families (Y axis).

Subsection 4. Construction of an evolutionary tree for prokaryotic species using information from many families

A species tree was built from the intergenome distance matrix derived with the three methods, incorporating information from all 1,379 protein families, using Neighbor Joining[88] as implemented in PHYLIP. Figure 11 shows the resulting tree. 1,000 bootstrap tree replicates were generated and used to evaluate the robustness of the topology. Bootstrap scores for the nodes with less than 95% confidence are shown. 49 (75%) of the 65 nodes in the tree have ≥ 95% confidence. Seven nodes have ≤ 50% confidence. Five of these seven nodes represent deep divergences, for example the 42% bootstrap support between the subtrees of Mollicutes, Firmicutes with Thermatoga.

Subsection 5. Comparison with species trees based on a small number of protein families

Comparison of the multifamily tree (Figure 11), with the previous 14 conserved family tree[37] and that in the Tree of life, based on thirty-one protein families with members in all included bacteria[35, 36], shows that the topologies of these trees are similar with the positions of two subgroups differing slightly. For example, the Bacillales subgroup (*Bacillus halodurans* and *Bacillus subtilis)* are adjacent in the multifamily tree (66% bootstrap score) and the tree of life, while they were grouped with two Listeria species in the 14 family tree (figure 11). We observe notable differences with respect to *Deinococus radiodurans*, *Thermotoga maritima* and *Aquifex aeolicus.* In the new tree *D. radiodurans* is grouped with the Actinobacteria

(93% bootstrap score), *T. Maritima* with Firmicutes (42% Bootstrap score) and *A. aeolicus* is grouped with Epsilon-proteobacteria (45% bootstrap score). In both trees based on a small number of families, these three organisms are grouped together. All are hyperthermophiles, and evidently that property dominates when only sequences from highly conserved protein families are considered, while a broader view suggests they are less closely related, although some of the relevant nodes are of lower confidence. Compared to the previous tree based on 14 highly conserved families, the branch lengths related to closely related species and strains are on average longer as a result of inclusion of faster changing sequences: The average fractional change in intergenome distance for 28 pairs of closely related species (those with distances less than 0.15 in the14 conserved protein family tree) in new multifamily tree compared to the 14 conserved family tree is +0.5293. (Seven pairs of genomes are still so close that the distances are not well resolved).

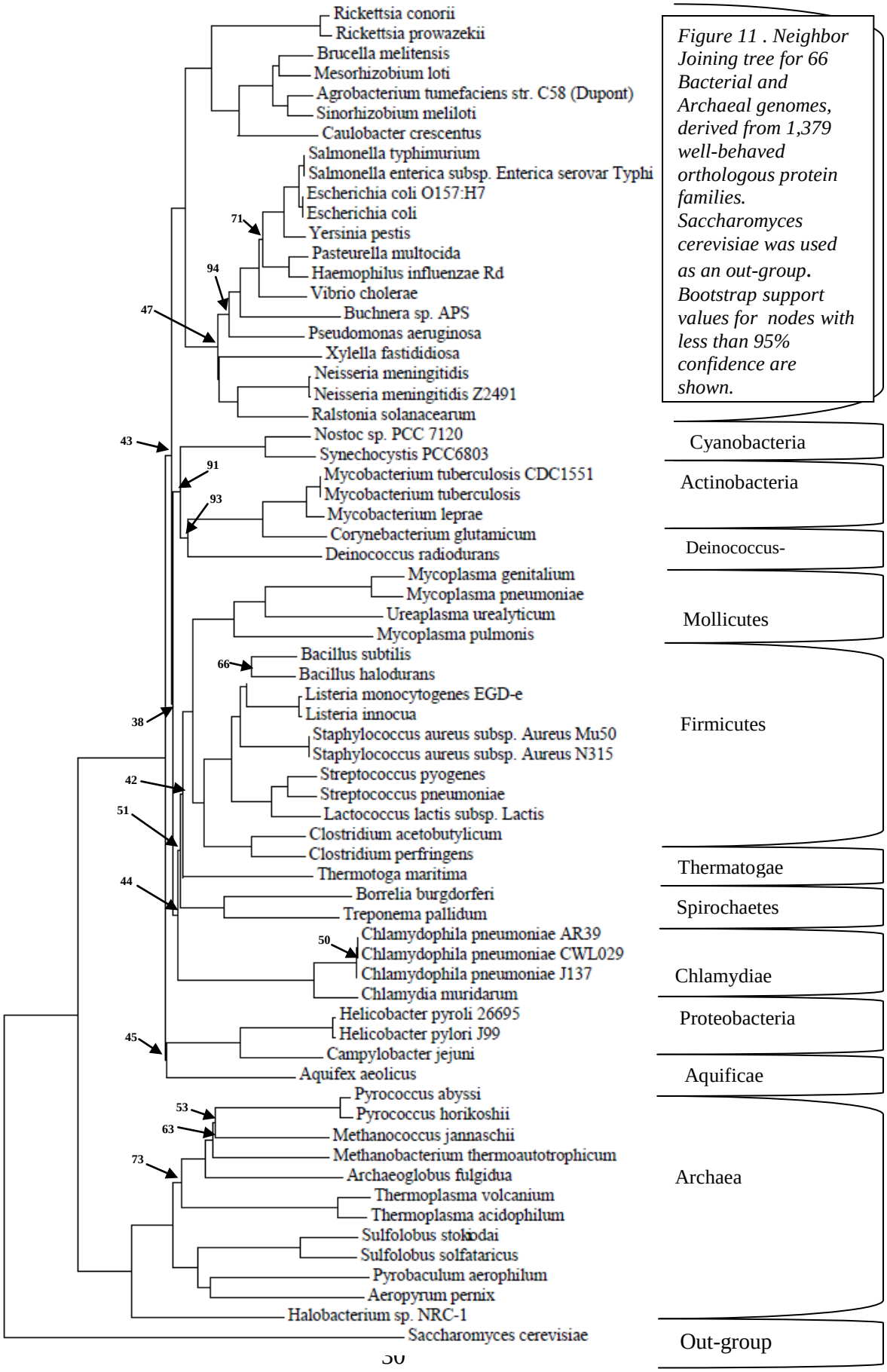


Figure 11 . Neighbor Joining tree for 66 Bacterial and Archaeal genomes, derived from 1,379 well-behaved orthologous protein families. Saccharomyces cerevisiae was used as an out-group. Bootstrap support values for nodes with less than 95% confidence are shown.

We also compared the multifamily tree with a more specialized Alphaproteobacteria tree[89] which was built using 104 families present in all 72 included species. The topology of the seven of these organisms included in our 66 genome set is identical with that in multifamily tree and the tree of life, but different from that in the 14 conserved family tree. For the archaeal kingdom, the multifamily tree topology is same as the 14 conserved protein family tree and as that of a thirty one conserved protein family archaea tree[90]. All 11 nodes of this kingdom have greater than 50% bootstrap confidence and eight out of 11 nodes have bootstrap scores greater than 95% confidence, indicating strong support for the topology.

Section 4 Discussion

Availability of many complete genome sequences provides new opportunities for reconstructing the relationship between species using a broad base of information. To exploit these opportunities, we have developed a method of utilizing information from a large number of protein families in constructing species trees. There are two major challenges to be overcome in incorporating the data from diverse protein families. First, as most families do not have members in all genomes and evolve at substantially different rates, a means of integrating the signals must be found. We address this by obtaining a relative evolutionary rate for each family, so providing a means of normalizing to a common evolutionary scale. Then, for each pair of genomes, we combine information from all the families with members in both to obtain an estimate of the evolutionary distance between that pair. Second, there are multiple sources of noise in the data that must be adequately contained. All gene level

phylogenetic reconstruction methods contend with problems of imperfect sequence alignment and the difficulties of reliably identifying appropriate orthologous relationships. We use state of the art methods for sequence alignment and orthologous family construction, but recognize these are imperfect. Additionally, for prokaryotes, analysis is greatly complicated by wide spread lateral gene transfer, for some families making the concept of linear descent from a common ancestor almost meaningless[26, 27, 28]. Our strategy for dealing with these three issues is two fold. First we use noise resistant methods for deriving relative evolutionary rates and intergenome distances, allowing robust determination of these quantities in many cases. Second, we identify families where the data are not consistent with linear descent by comparing the results from three methods, and discard the inconsistent families. The availability of information from thousands of families makes this approach practical. As a result, we obtain intergenome distances based on information from 1,379 orthologous families with evolutionary rates varying by an order of magnitude and orthologous family members present in from four to 66 genomes.

A disadvantage of the approach is that the large numbers of sequences involved preclude the use of character based descriptions and associated optimization methods[86]. On the other hand, the new method is scalable to the inclusion of very large numbers of genomes. Inclusion of information from a large number of diverse families also allows a broader and less biased view of the differences between species.

All methods for deriving these evolutionary relationships assume that the values of some feature or features are correlated with speciation and time. Before the

availability of molecular and genome level data, these features were often morphological. Morphological features may vary under environmental and other selective pressures in a manner that is not directly correlated with speciation processes, for example converging to similar values for bacteria and archaea[91]. Non-speciation related variation of molecular properties also occurs, for example, repeated switching back and forth of enzyme specificity within orthologous families, as in the case of malate and lactate dehydrogenase[92]. Sequence based phylogenetic methods assume that overall sequence identity relationships within orthologous families are not substantially distorted by these sorts of effects. While generally true, there may be exceptions. For example, adaption to a specific environmental condition, such as temperature, may cause selection of particular amino acid types, and so constrain sequence similarity in a manner not directly related to speciation. We see evidence of that among the bacterial hyperthermophiles. The three hyperthermophylic species included in our analysis grouped in the same sub-tree in two previous phylogenetic analyses using a small number of protein families[34, 35, 37], but are in three separate sub-trees with the larger number of families used in this work. That result suggests highly conserved families may exhibit temperature correlated sequence similarities, or simply that a small sample of families is more likely to be unrepresentative of the time course than a large number.

Particular processes may change more rapidly in some periods than others, resulting in atypical rates of sequences change for the proteins involved. For example, archaeal and bacterial ribosomes are markedly different in overall structure and composition[81, 82], and that appears to be reflected in relatively large sequence

differences between their orthologous proteins. Because there are so few protein families with members in all genomes, methods that rely on that feature include a large fraction of ribosomal proteins, for example 21 of the 31 families in the Tree of Life[27, 32, 33, 34, 35, 35] are ribosomal. Thus, our previous 14 family tree shows larger distances between the archaeal and bacterial kingdoms than the new many family tree, likely reflecting the bias introduced by the ribosomal proteins. The new tree also has better resolved branch lengths for closely related species and strains, reflecting the fact that inclusion of families with faster changing sequences provides a better numerical basis for determining these. Other differences between trees built with a small versus a large number of protein families, such as *B.subtilis* and *B.Halodurans* are adjacent in the new multifamily tree and similar to the tree of life, are less easily traced to specific effects. In general, though, the more families included the less impact from effects that are not closely coupled to speciation.

A wide choice of families to include also opens up the possibility of examining the rate of evolution and adaptation of particular processes and functions – all families involved in a particular GO[93] process, such as cell division, might be considered, for example.

Section 5 Materials and Methods

Subsection 1.Orthologous protein domain families

The work utilized a set of 31,874 protein domain families previously compiled from the complete genome sequences of 66 representative prokaryotic genomes[8]. These

families include 20,992 singletons (families with only one member), 4,810 doubletons and 6,072 protein families containing three or more members. 4,856 primary orthologous families were extracted from the 6,072 domain protein families with three or more members[38]. All analysis was performed on this orthologous set.

Subsection 2. Calculation of the average accepted amino acid substitutions per site between each pair of domains '*i*' and '*j*' in each orthologous family '*u*', *S*(*i*,*j*,*u*)

Multiple sequence alignments for each family were generated using MUSCLE[94]. The maximum likelihood average accepted amino acid substitutions per site between each pair of domains '*i*' and '*j*' in each family '*u*', *S*(*i*,*j*,*u*), were obtained from these alignments using the PROTDIST module in PHYLIP[95] with the Jones-Taylor-Thornton (JTT) amino acid substitution matrix[96].

Subsection 3. Initial intergenome distances derived from a set of 14 highly conserved families

Initial intergenome distances between all pairs of the 66 genomes were derived using a set of 14 conserved orthologous protein families, all with members in each genome. 12 of these are the ribosomal proteins (L2, L5, L10, L13, L14, L15, S2, S3, S5, S11, S13, and S17). The other two protein families are the DNA-directed RNA polymerase (alpha subunit) and the Preprotein translocase secY subunit. Intergenome distances, $S_{14}(i,j)$, between genomes *'i'* and *'j'* were calculated by averaging accepted amino

acid substitutions per site between each pair of domains '*i*' and '*j*' in each of the 14 families '*u*', *S*(*i*,*j*,*u*).

$$S_{14}(i,j) = < S(i,j,u) >_u$$

Subsection 4. Calculation of relative evolutionary rates for each orthologous protein sub-family

Within a family 'u', the relative rate of sequence change between any pair of genomes *'i'* and *'j'* is expressed as:

$$r(i,j,u) = S(i,j,u) / S_{14}(i,j)$$

Averaging over all pairs of genomes with members in the family then provides an estimate of the rate of sequence change, relative to the rate for the conserved families, *R(u)*.

$$R(u) = <r(i,j,u)>_{i,j}$$

r(i,j,u) values are noisy because of the effect of lateral gene transfer, possible errors in sequence alignment, and possible errors in identification of orthologous relationships, so that a straight average of this form is unreliable. We use three different methods designed to handle such noisy data: least median squares[97, 98], a Gaussian kernel density estimator[99], and a recursive filtering method.

1. Least median squares (LMS)[97, 98] was used to find the value $R_{LMS}(u)$ with the minimum median square value of the residual set, $\{\delta(i,j,u)^2\}$ where

$$\delta(i,j,u) = r(i,j,u) - R_{LMS}(u)$$

and the set includes contributions from all pairs of genomes '*i*' and '*j*' with members in family '*u*'. Median squares are much less sensitive to outliers than the more usual least squares procedure. Formally, for conventional least squares the breakdown point (the smallest fraction of contamination that can falsify the linear estimator, where "falsify" is defined as changing the regression line by 90 degrees) is *1/n*, where *n* is the number of data. For median squares the breakdown point is 50%, the highest breakdown point theoretically possible[100]. Inspection of the value of $R_{LMS}(u)$ for many families shows effective robustness to obvious outliers for the family rate data.

2. A Gaussian kernel density estimator (GKDE)[99, 101] was used to represent the probability density of *R(u)* as a sum of gaussians, one gaussian centred at each value of *'r'*. For each family *'u'*, Gaussian kernel density distributions were compiled with the KernSmooth in R[101] using the set of *{r(i,j,u)} values* for all *'n'* pairs of genomes *'i'* and *'j'* containing members of the family. The total rate density $\rho(r')$ at any value of *'r'* is:

$$\rho(r') = \sum_{k=1}^{n} e^{\frac{-(r'-r_i')^2}{h^2}}$$

h is the bandwidth set equal to $\frac{0.9\hat{\sigma}}{n^5}$, where $\hat{\sigma} = min(s, Q/1.34)$, $s^2 = \frac{1}{n-1}\sum_i^n (r_i - \bar{r})^2$ and *Q* is the interquantile range of the data[102]. The maximum value of ρ corresponds

to the maximum empirical likelihood value of $r'(u)$, taken to be the GKDE estimate of the relative evolutionary rate for this family ($R_G(u)$).

3. A simple recursion filtering procedure (RFP) was used to iteratively estimate the value of the relative rate of sequence change for family u, $(R_{RF}(u))$, as $<r(u)>_n$ rejecting outliers (those differing by more than 0.5 substitutions per site from the current average) at each iteration. In practice, this procedure converges after three iterations.

An initial combined estimate of the relative evolutionary rate $R_{AVG}(u)$ for each family *'u'* was obtained by averaging over the values obtained by the three methods. The subset of families with consistent rates across the methods (rates differing by 20% or less), that is:

$$|R_{LMS}(u) - R_G(u)| \leq 0.2R_{AVG}(u) \text{ and } |R_{LMS}(u) - R_{RF}(u)| \leq 0.2R_{AVG}(u)$$

was used for subsequent analysis.

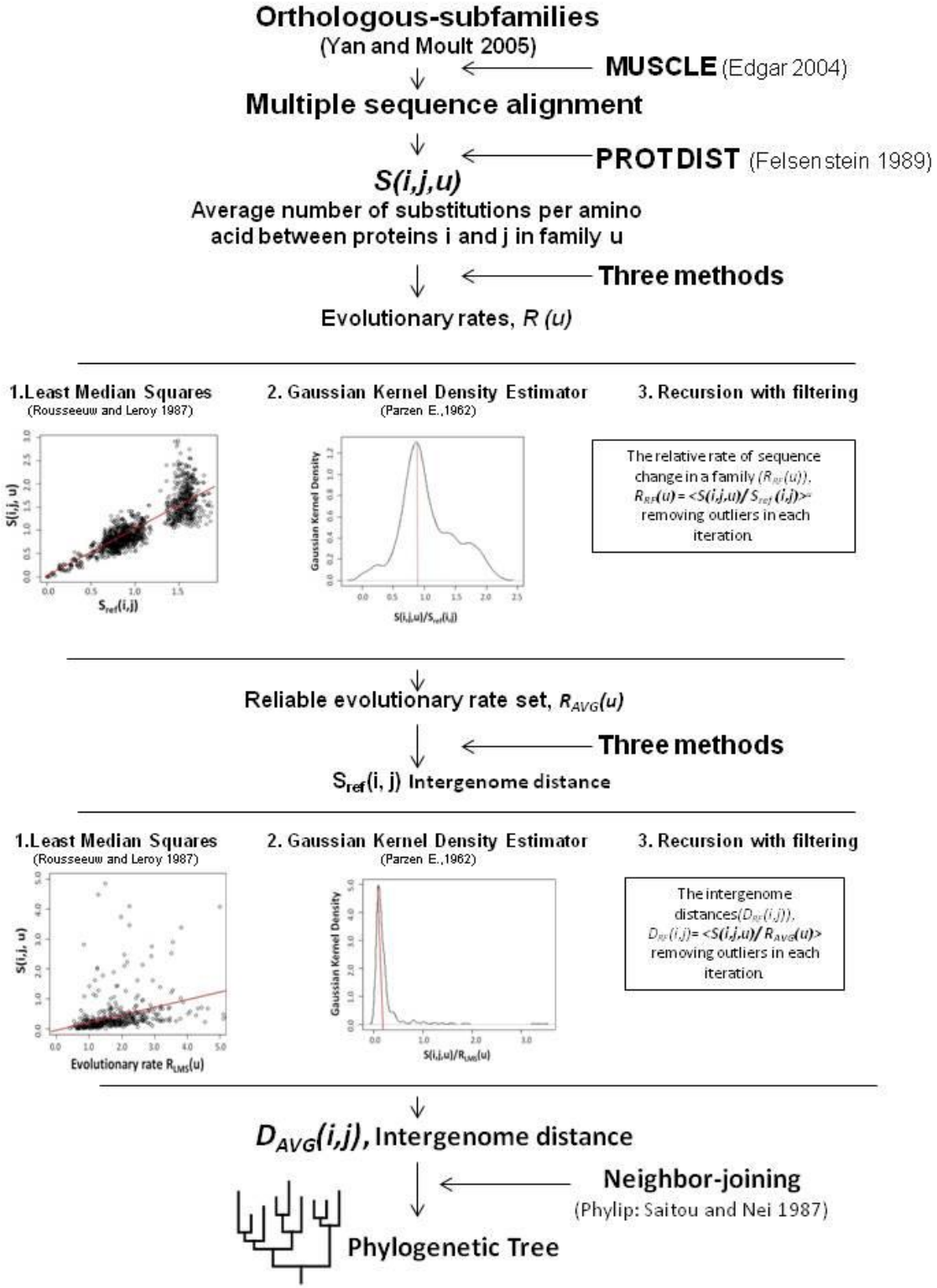

Orthologous-subfamilies
(Yan and Moult 2005)
MUSCLE (Edgar 2004)
Multiple sequence alignment
PROTDIST (Felsenstein 1989)
S(i,j,u)
Average number of substitutions per amino acid between proteins i and j in family u
Three methods
Evolutionary rates, R (u)
1.Least Median Squares
(Rousseeuw and Leroy 1987)
2. Gaussian Kernel Density Estimator
(Parzen E.,1962)
3. Recursion with filtering
The relative rate of sequence change in a family ($R_{RF}(u)$), $R_{RF}(u) = \langle S(i,j,u) / S_{ref}(i,j) \rangle^{-\alpha}$ removing outliers in each iteration
Reliable evolutionary rate set, $R_{AVG}(u)$
Three methods
$S_{ref}(i, j)$ Intergenome distance
1.Least Median Squares
(Rousseeuw and Leroy 1987)
2. Gaussian Kernel Density Estimator
(Parzen E.,1962)
3. Recursion with filtering
The intergenome distances($D_{RF}(i,j)$), $D_{RF}(i,j) = \langle S(i,j,u) / R_{AVG}(u) \rangle$ removing outliers in each iteration
$D_{AVG}(i,j)$, Intergenome distance
Neighbor-joining
(Phylip: Saitou and Nei 1987)
Phylogenetic Tree

Figure 12. Flowchart of the procedure used to estimate the evolutionary rates of orthologous protein families and Intergenome distances. Three methods are used: least median squares, a Gaussian kernel density estimator, and a recursive filtering method. The final set of intergenome distances was used to reconstruct a phylogenetic tree.

Subsection 5. Estimation of intergenome distances using information from many protein families

Each member of the set of $S(i,j,u)$ values with members in genomes '*i*' and '*j*' provides information concerning the relative intergenome distance $D(i,j)$ between those species. In order to combine the information from all contributing families, the intergenome distances $S(i,j,u)$ are placed on the same scale by normalizing with the relative evolutionary rate for that family:

$$S'(i,j,u) = S(i,j,u) / R_{AVG}(u)$$

Information from the set of $S'(i,j,u)$ values for a pair of genomes *'i'* and *'j'* is combined to provide an estimate of the relative intergenome distance, $D(i,j)$. To combat noise in the $S(i,j,u)$ values we make use of the same three robust methods described above. For least median squares, the value of $D_{LMS}(i,j)$ with the minimum median square value of the residual set, $\{\delta(i,j,u)^2\}$ is found, where

$$\delta(i,j,u) = S'(i,j,u) - D_{LMS}(i,j)$$

and the set includes contributions from all families that have members in genomes '*i*' and '*j*' that have consistent evolutionary rates. For the Gaussian kernel density

estimator[99], Gaussian kernel density distributions were compiled for the set of $S'(i,j,u)$ values for all families with members in genomes *'i'* and *'j'* that have consistent evolutionary rates and the value of S' with maximum density $\rho(S')$ taken as the maximum empirical likelihood of value the intergenome distance $D_G(i, j)$. For recursive filtering, three rounds were performed, rejecting contributions in the second and third rounds for which

$$|D_{RF}(i,j)) - S'(i,j,u)| > 0.5$$

A final combined estimate of the intergenome distance $D_{AVG}(i,j)$ for each pair of genomes *'i'* and *'j'* was obtained by averaging over the values obtained by the three methods. For those pairs of genomes with consistent intergenome distances across the methods, that is where:

$$|D_{LMS}(i,j) - D_G(i,j) \leq 0.2\ D_{AVG}(i,j) \text{ and } |D_{LMS}(i,j) - S_{RF}(i,j)| \leq 0.2\ D_{AVG}(i,j)$$

$D_{AVG}(i,j)$ values were used to provide the elements of the distance matrix for reconstructing a species tree. For genome pairs where inconsistent distances were obtained, $D_{LMS}(i,j)$, judged to be the most reliable single method, was used.

An example of determination of a family evolutionary rate using least median squares, for the mercuric resistance operon repressor protein (merR) family, is shown in Figure 13. This family has members in 39 genomes, providing amino acid substitution values from 1,482 pairs of genomes, each contributing to the determination of its relative evolutionary rate. There is a substantial scatter of

contributions to *S(i,j,u)* values contributed by different genome pairs, but a group of points, colored green, and all having one member in *Archaeoglobus fulgidus (aful)* are clearly separated from the rest, consistent with the member in that genome being the result of lateral gene transfer. Least median squares produce the red line, successfully ignoring these outliers.

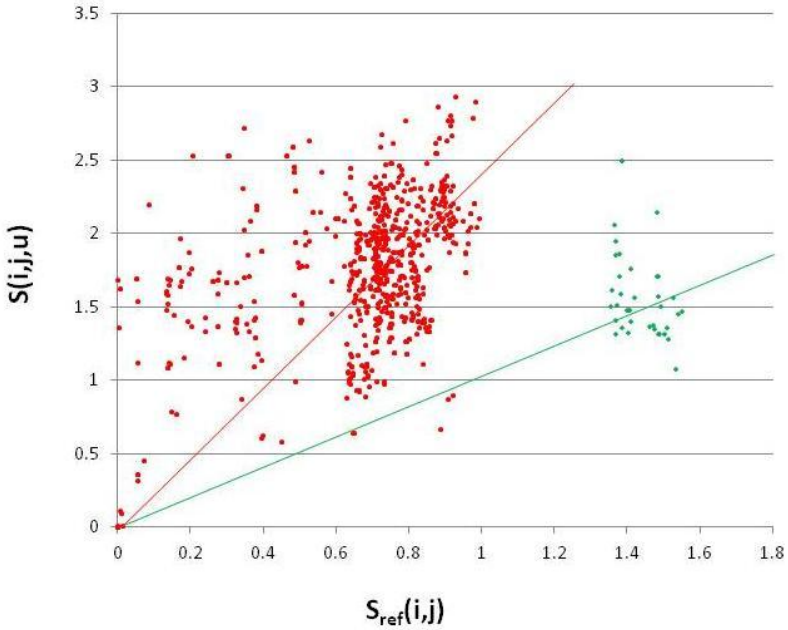


Figure 13. Example of determining the relative evolutionary rate for the *mercuric resistance operon repressor protein (merR) family using least median squares (LMS). The LMS line is shown in red, and the slope (2.49) gives the relative rate,* $R_{LMS}(u)$*. Green points are for intergenome distances apparently involving a lateral gene transfer event. The LMS fit (red line) effectively ignores these and other outliers.*

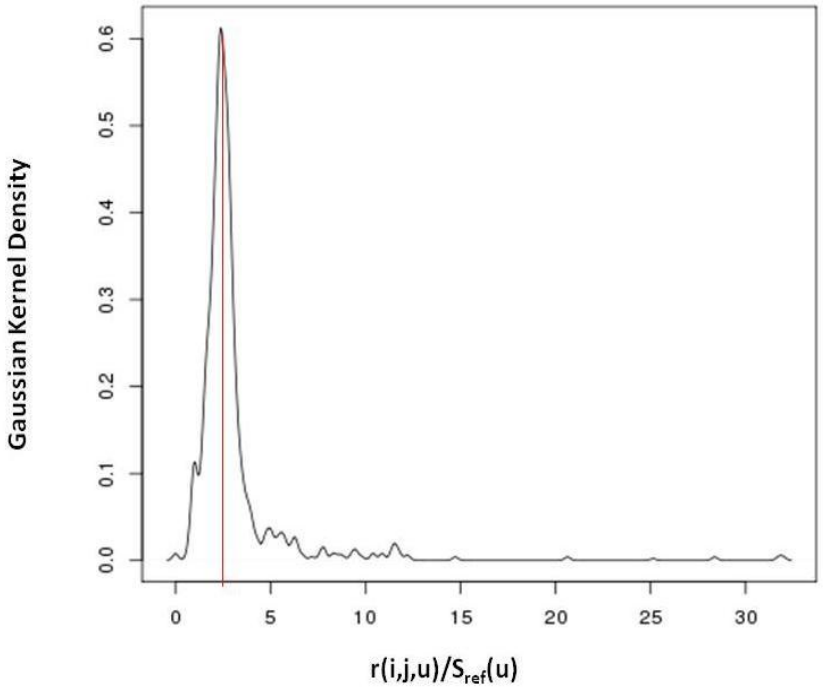


Figure 14. Example of determining the relative evolutionary rate for mercuric resistance operon repressor protein (merR) family using a Gaussian kernel density estimator. The red line indicates the highest density, taken to be the relative evolutionary rate for the family, $R_G(u)$ at 2.77. Outliers caused by LGT contribute the small sub-peak at ~1.0. Other outliers form a high value tail to the distribution.

Figure 14 shows the corresponding Gaussian kernel density distribution, with the LGT affected points forming a small sub-peak at about ~1.0, and other outliers with anomalously high values producing a tail from the main peak. The peak has a value of 2.77, close to that obtained with the LMS analysis.

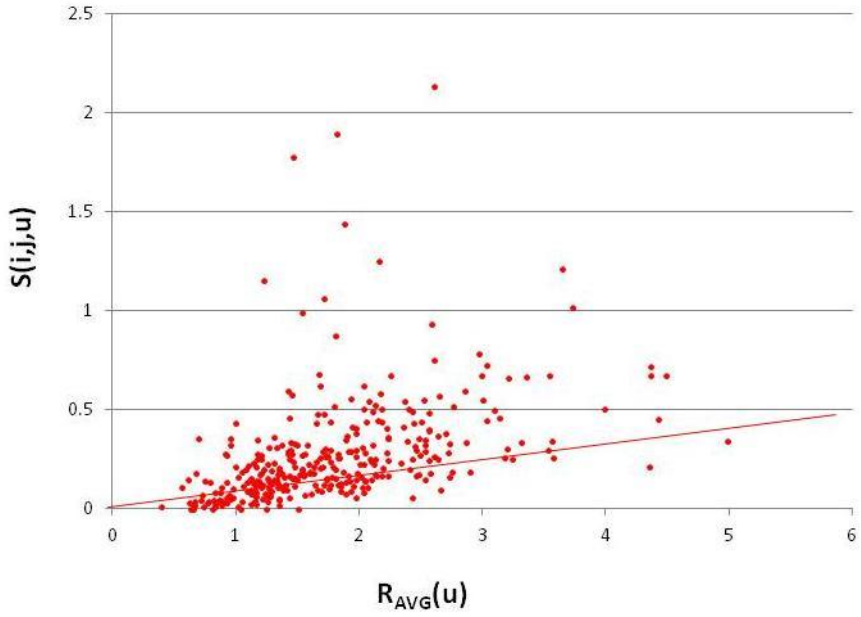


Figure 15. Example of determining the intergenome distance between a pair of species (Haemophilus influenzae and Pasteurella multocida) using least median squares (LMS). The LMS line is shown in red, and the slope (0.11) gives the intergenome distance, $D_{LMS}(i,j)$. The anomalously large s(I,j,u) values have a little effect on the derived distance.

An example of estimation of an intergenome distance using least median squares. $D_{LMS}(i,j)$, is shown in figure 15. 330 families with consistent evolutionary rate estimates have members in these two genomes and so contribute to determining this distance. Figure 15 shows there are a number of outliers, effectively ignored by the LMS fit.

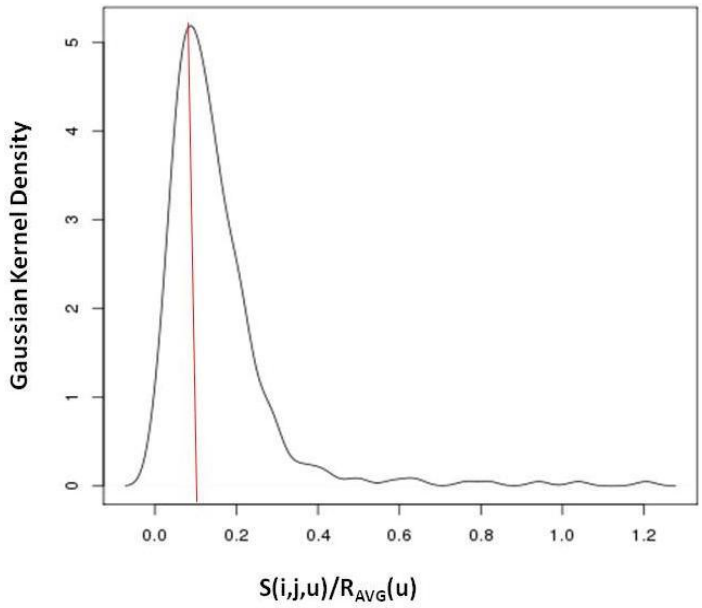


Figure 16. Example of determining the intergenome distance between two species (Haemophilus influenzae and Pasteurella multocida) using a Gaussian kernel density estimator. The red line indicates the highest density, at 0.09, taken to be the estimated intergenome distance for this pair of genomes, $D_G(i,j)$. The anomalously high values in the distribution result in an asymmetric peak, but have no effect on the maximum.

Figure 16 shows the Gaussian kernel estimator result for the same pair of genomes. Here the outliers form a tail from the main peak, not significantly influencing the peak value. In this case, the three methods all return similar values: least median squares, 0.11, Gausssin kernel estimator, 0.09 and recursive filtering, 0.12

Subsection 6. Construction of a species tree based on multi-family intergenome distances

Phylogenetic trees were built from the matrix of intergenome distances using the Neighbor Joining method[88] as implemented in PHYLIP. The tree topology robustness was evaluated using a bootstrap procedure[103]. 1,000 trees were built. For each tree, N families were randomly selected, with repetition, where N the number of families included. For each selected set of families, intergenome distances were re-determined as described above. These distances were then used to build a tree. A Consensus Tree procedure[95] was used to compare the topology of these thousand trees and to obtain confidence values for each subtree.

Chapter 3: Molecular evolution of protein families: The properties of proteins as a function of age

Section 1Abstract

Phylogenetic analysis of sets of complete genomes has revealed that most protein families appear to have recently emerged. One hypothesis for the origin of these families is that they represent new open reading frames that have been created either from previously noncoding DNA, by frame-shifting from older open-reading frames, or are the result of recombination of sub-domain fragments from older proteins. A test of this hypothesis is whether or not proteins in young families have substantially different properties from those in older families, consistent with a recent origin. To this end, we have examined four properties of protein families to determine whether or not there is evidence to support the new open reading frame hypothesis.

Methods: A set of 66 prokaryotic genomes is used for the analysis. Orthologous protein family age was estimated from the phylogenetic distribution of family members in a previously compiled species tree. Age distortion arising from lateral gene transfer was reduced by removing proteins with anomalous rates of sequence change. Four quantities were considered as a function of family age: mRNA expression level, relative rate of change of amino acid sequence within each family, level of predicted intrinsic structural disorder, and the number of known protein-protein interactions. A partial correlation analysis was used to control for interaction between variables.

Results: A strong correlation was found between each of the four quantities considered and the apparent age of the families. The partial correlation analysis results are consistent with age as the driving variable for all four. Average expression level increases 16 fold between the youngest and the oldest families; average evolutionary rate is five times slower for the oldest families than for the youngest, and the average number of protein partners is five times as large for the oldest families as for the youngest. Average predicted structural disorder also decreases with age, reaching a level two times lower than that of the youngest families, before rising slightly for the oldest subset of families. All these observations are consistent with structural and functional immaturity for the majority of proteins in young families, and thus consistent with recent origins of their open reading frames. An interesting additional observation is that the apparent correlation between E.coli K12 mRNA expression and family evolutionary rate, noted by others for this and several additional species, is an artifact of both these quantities correlating with age. Thus the often proposed explanation that the expression/rate correlation arises as a result of negative selection of variants in highly expressed proteins may not be correct.

Section 2 Introduction

Prior to the advent of the first fully sequenced genomes the prevailing model of protein evolution was that all proteins have descended from a relatively few ancient ancestors. That is, all protein families are old on an evolutionary time scale, and that there are of the order of only 1000 independent evolutionary lines[1]. Once complete sequences were available for a number of genomes, it became clear that the data are not consistent with this model, with a substantial fraction of open reading frames in each genome apparently unrelated to any previously sequenced proteins (so called singletons or Orphans)[6]. As more genomes have been sequenced, this picture has refined into a view that many protein families are phylogenetically narrow in distribution, in a manner consistent with a relatively recent origin[104]. There are a number of possible origins of these apparently young families: (1) these proteins may in fact belong to ancient families, but have diverged sufficiently fast in sequence that sequence relationships are not powerful enough to detect relatives. Two lines of evidence suggest this is not the case. First, as more and more genomes are sequenced, most of these families remain apparently young. Second, an analysis of protein structures suggests that by this more sensitive measure there are a large number of independent evolutionary lines[4]. (2) In prokaryotes, these proteins may belong to ancient families populating so far unexplored phylogenetic regions, and have recently undergone lateral gene transfer (LGT) to their present relatively isolated locations. Although LGT within the prokaryotic kingdom is very common[29, 30, 31], the addition of many more prokaryotic genomes (now over 10,000 (http://www.ncbi.nlm.nih.gov/genome/browse/) has not revealed extensive origins of this type. A significant fraction of apparently young families (about 25% by

one estimate[25] do appear to have transferred from phages, but that does affect the apparent age distribution significantly. (3) These proteins are in some sense really new, either created by frame-shifting of a previous open reading frame, or occurring in previously non-coding DNA, or arising from recombination of fragments from two or more older proteins. Recombination of domains to form new multi-domain proteins is very common, especially in Eukaryotes[105], but evidence of sub-domain recombination is rare. Experimentally, proteins belonging to apparently young families have proven difficult to study, with low success in purification and crystallization for X-ray studies (for example, Vitkup D. et al paper[106]), suggesting less robust structural and stability properties than most proteins, as might be expected for recently established new open reading frames .

In this work, we have investigated the hypothesis that these apparently young proteins are immature by examining a number of properties as a function of the apparent age of the corresponding orthologous family. Specifically, we investigate rates of sequence change within families, mRNA expression levels, amount of structural disorder, and complexity from a functional standpoint, as monitored by the number of interactions with other proteins. We examine the level of each of these properties as a function of the apparent age of the orthologous protein family concerned. 'Age' is defined in terms of the phylogenetic distribution of members of the family in a species tree. The tree was previously built using information from many protein families, and so is expected to have more robustly determined branch lengths than trees based on data for just a few families. We use a set of 66 prokaryotic genomes for this analysis, and reduce the

distortion of apparent age arising from lateral gene transfer by omitting proteins most likely to have been involved in that process.

Section 3 Results

Subsection 1. Orthologous protein domain families

As described in methods (Section 5), the analysis was performed on the set of 1,196 primary orthologous families from 66 prokaryotic genomes that have a member from E.coli K12[8]. Of these families, 94, 151 and 951 are singletons (families with only one member), doubletons (families with two members) and multitons (families containing three or more members) respectively.

Subsection 2. Family age

The relative age of each protein family was derived from the species tree, as described in Methods (Section 5). Figure 17 shows the distribution of relative family age for the 1,196 orthologous families. Ages are in units of accepted substitutions per site in the reference set of 14 highly conserved slowly evolving families (Chapter 2) and are calculated after removing proteins most likely involved in lateral gene transfer events (see Section 5 Methods). Ages range from 0.00158 for singletons of Escherichia coli K12 to 0.978 for families with apparent origins at the base of the tree.

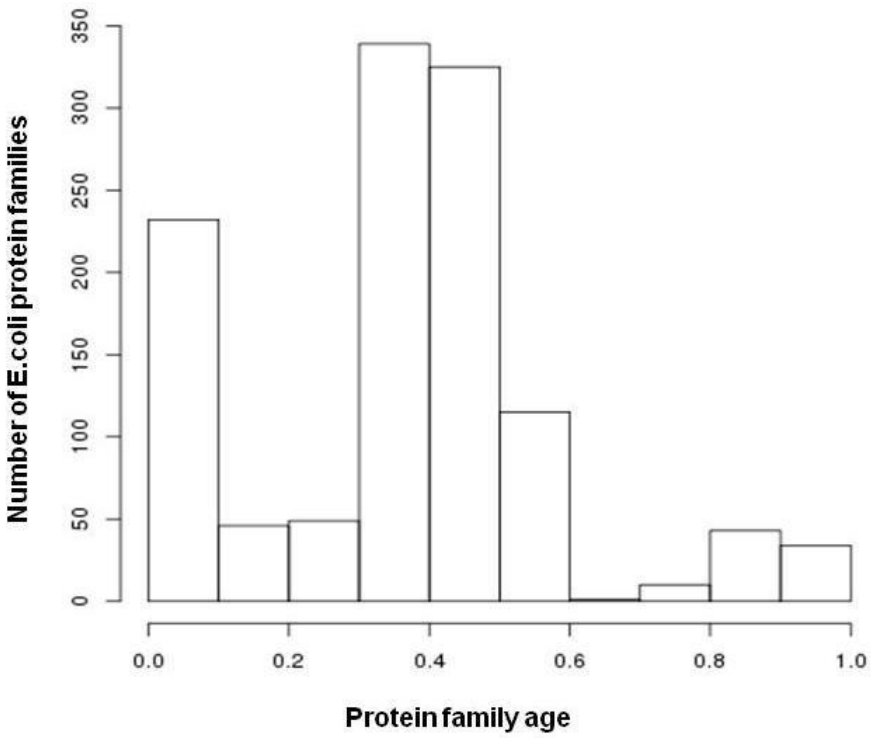


Figure 17. Distribution of relative family ages. Age is in units of accepted substitutions per site in a reference set of 14 conserved families.

Subsection 3. Relationship between Expression level and apparent family age

Figure 18 shows the distribution of RNA expression level for the E.coli members of the 971 of the 1,196 E.coli proteins for which measurements are available. Red bars show the distribution for the family age less than and equal to 0.15 (all are singletons and doubletons with family ages ranging from 0.00158 to 0.152. Blue bars show the distribution for remainder. Relative $\log_2$ mRNA expression levels vary widely, ranging from -7.97 to 5.76. Although the average expression level for the youngest proteins is lower than the older ones, most do have clearly detectable expression, providing confidence that they are in fact Bona Vida open reading frames in the genome.

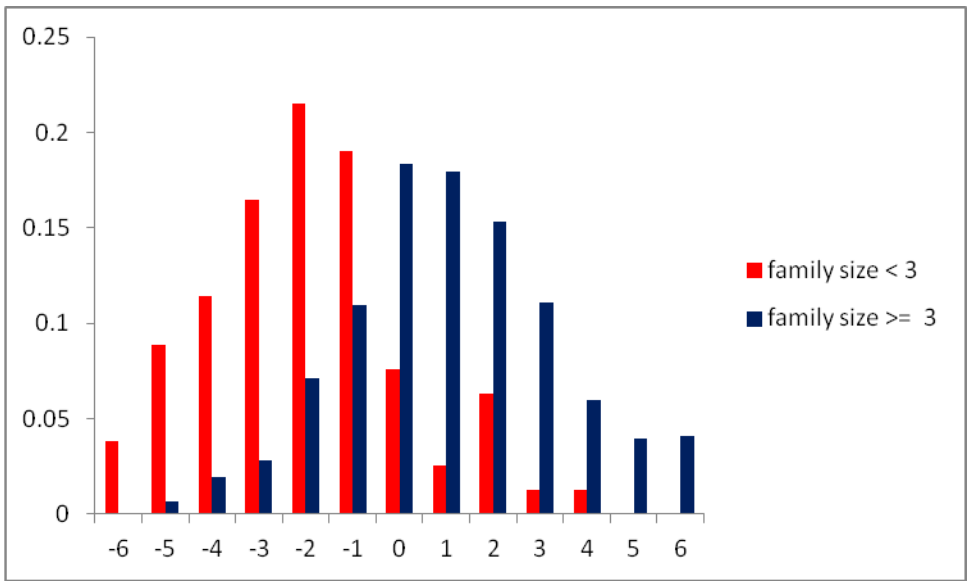
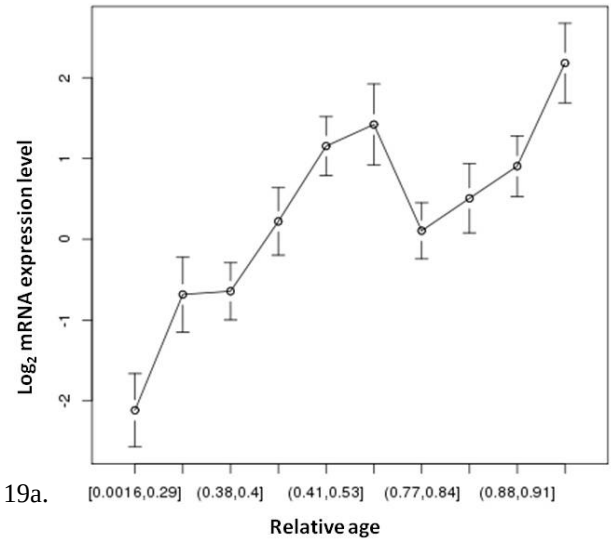
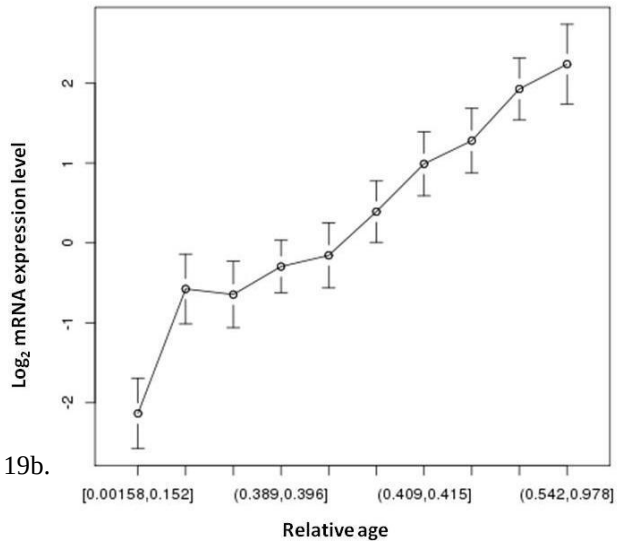


Figure 18. Distribution of expression levels for 971 E.coli proteins. Red bars show the distribution for singletons and doubletons, and blue bars show the distribution for multitons. Y-axis is the fraction of E.coli proteins in each expression level bin (expression data are from http://www.genome.wisc.edu/, log phase growth on glucose). X- axis shows log_2 relative mRNA level.



19a.



19b.

Figure 19. Improvement in estimates of protein family age by partial removal of Lateral Gene Transfer (LGT) events. (a) Comparison of average log2 mRNA expression level as a function of apparent family age for 971 E.coli proteins in the orthologous subfamilies, and (b) the same, omitting the 15% of proteins most likely to have undergone LGT. There is a steady increase in mRNA level with age (expression data are from http://www.genome.wisc.edu/, log phase growth on glucose). Y-axis shows average of log_2 relative mRNA level, bars show 95% confidence intervals

(approximately 2 sigma). Equal points per age bin, 'age' in units of average accepted substitutions per site in a set of conserved protein families.

Figure 19 shows the comparison of E.coli K12 mRNA expression with family age, including all proteins in the 971 families with E.coli members and expression measurements. There is a very strong correlation between apparent family age and E.coli K12 expression level (P values $< 2.2e-16$ by both Pearson Correlation and Kendall Tau). There is a steady increase in average expression level with family age, with average expression level of the oldest and youngest families differing by a factor of about 16.

Subsection 4. Relationship between relative rates of amino acid change and apparent family age

The relative rates of evolutionary change of each family, as reflected in the rates of amino acid change, were taken from our previous analysis (Chapter 2). Figure 20 shows the distribution of these rates for the 514 out of the 1,196 primary orthologous families for which reliable rates were obtained. Most rates lie in the range from 0.39 to 10, with a long tail up to 30.5 times that of the average rate of the reference 14 conserved protein families.

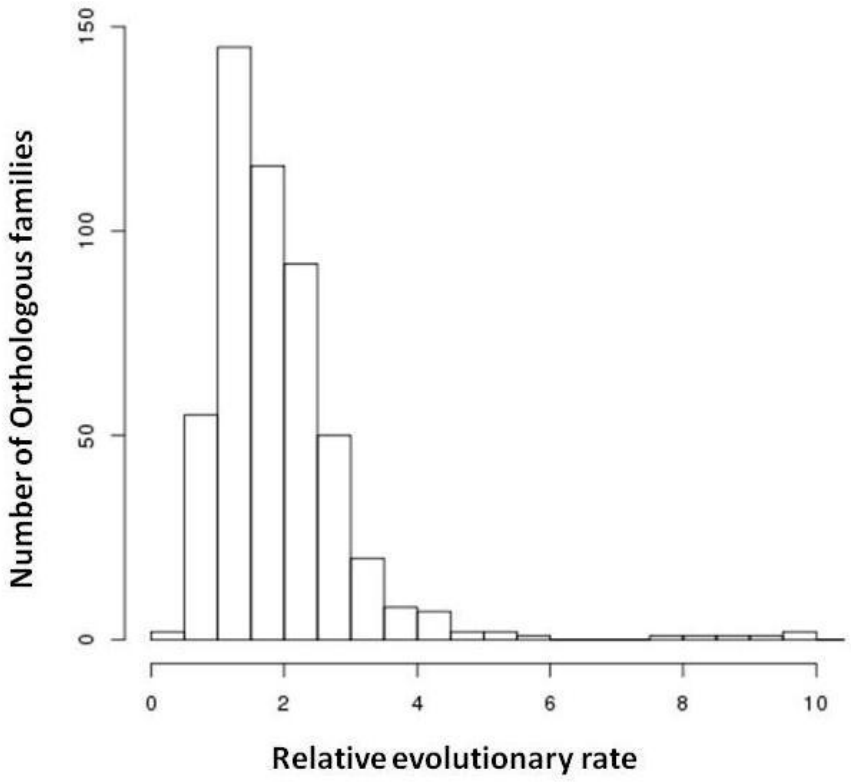


Figure 20. Distribution of relative evolutionary rates for a set of 514 orthologous protein families. 1.0 corresponds to the average rate of amino acid substitutions for 14 highly conserved families that have members in all 66 genomes considered.

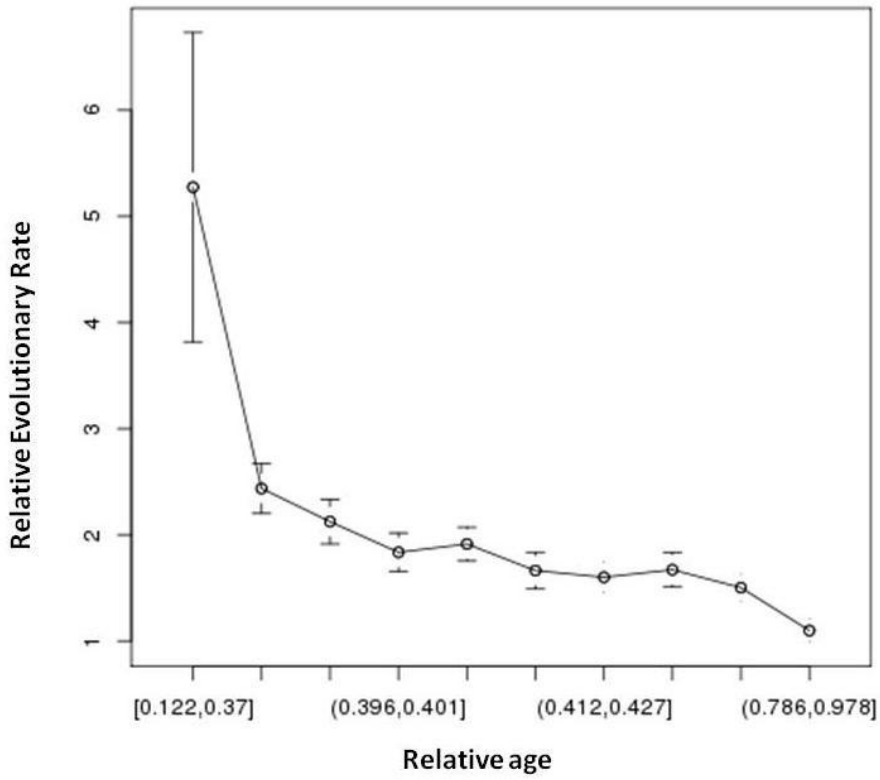


Figure 21. Average relative evolutionary rates in 514 orthologous families as a function of apparent age. 'Age' in units of average accepted substitutions per site in 14 conserved families, rate relative to the average rate of sequence change in those families. Rates in the youngest set of families are more than twice the overall average, and there is a decrease in average rate of approximately two fold over the remainder of the age range. (Bars show 95% confidence intervals, equal points per bin).

Figure 21 shows the comparison of relative evolutionary rates with family age. There is an obvious strong negative correlation between these variables (P-value < 2.2e-16 by both Pearson Correlation and Kendall Tau), and the largest rates observed to the youngest families, and an overall five-fold spread in average rates.

Subsection 5. Relationship between the number of protein-protein interactions and apparent family age

Many proteins are involved in interactions with other proteins, and we are interested in the extent to which the number of such interactions evolves with age. For this purpose, we retrieved a curated set of experimentally determined E.coli K12 protein-protein interactions[107] (see Section 5 methods). The distribution of the known number of protein-protein interaction partners ranges from 0 to 33 (Figure 22). Although this type of data is plagued by high false positive and false negative rates, we do not expect the extent of these errors to correlate with the other properties, such as age, that we are interested in.

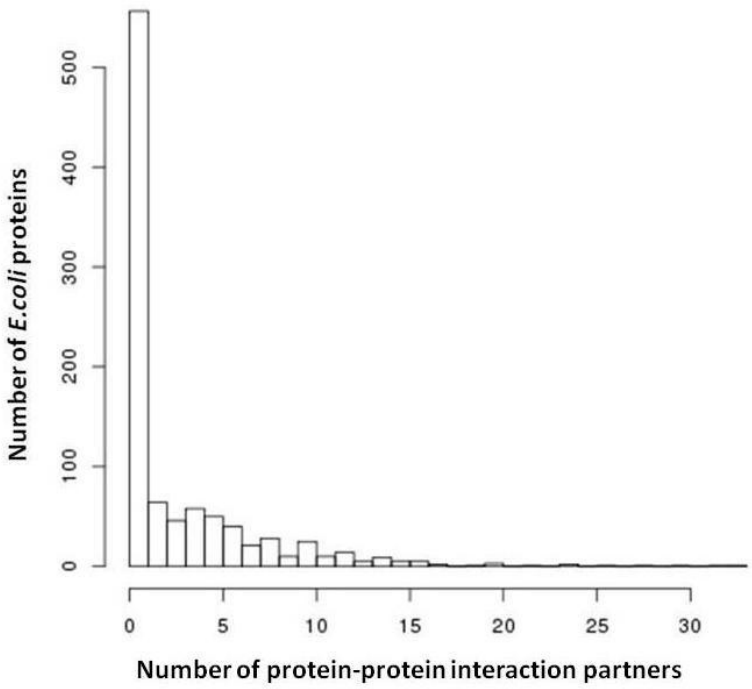


Figure 22. Distribution of the number of known protein-protein interaction partners for 1,196 E.coli proteins[98].

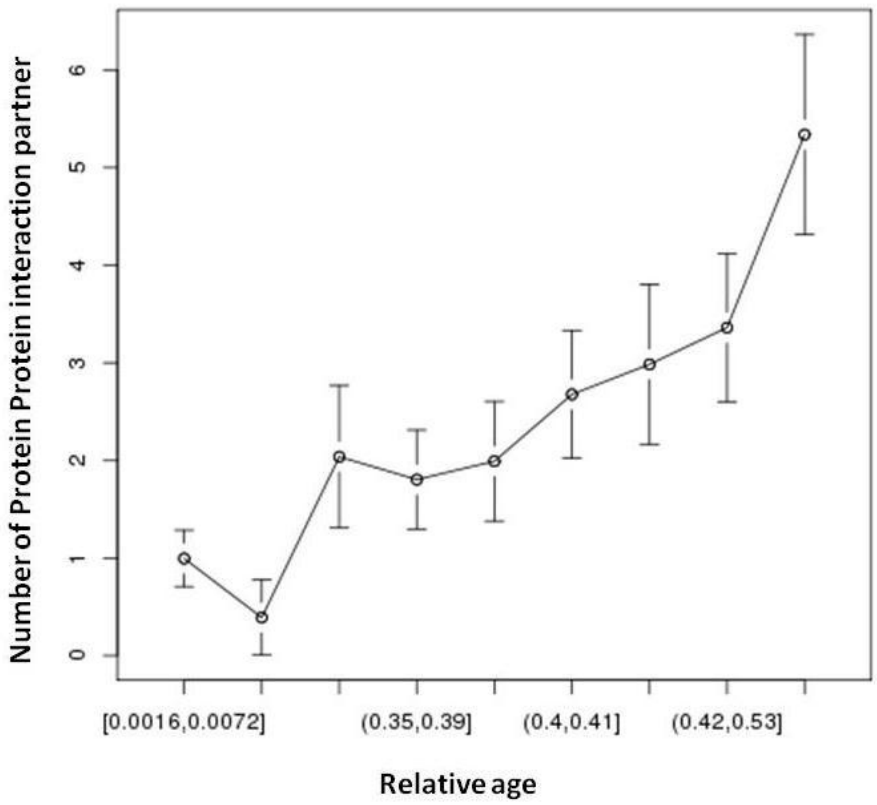


Figure 23. N*umber of known protein interaction partners in 1,196 Escherichia coli proteins as a function of the apparent age of the corresponding families. There is increasing the number of interaction partners with increasing family age. (Bars show 95% confidence intervals, equal points per bin.*

Figure 23 shows that *E.coli* K12 proteins in young families have the fewest known interaction partners (with an average of about 1). The number of partners steadily increases with age, so that the proteins in the oldest families have an average of more than 5 times as many partners (P-value = 6.45e-07 by Pearson correlation and P-value < 2.2 e-16 by Kendall Tau correlation). This result is consistent with the idea that the longer a protein is present in the interactome, the more partners it acquires. However, as always, correlation is not cause, and we will see later that other explanations are possible.

Subsection 6. Relationship between intrinsic disorder and apparent family age

To address the question of variation of the amount of intrinsic disorder in proteins as a function of family age, we calculated predicted percentage of disordered residues in all included 1,196 E.coli protein sequences using DISOPRED2[73]. Figure 24 shows the distribution of % predicted disorder. Contrary to an earlier study of E.coil proteins[108], the large majority of proteins have less than 20% of disorder, and very few have more than 50%. These results differ from an earlier study that predicted 50% of E.coli proteins to have greater than 40% disorder. We attribute this difference to the tendency of the alternative method[109] to over-predict[110].

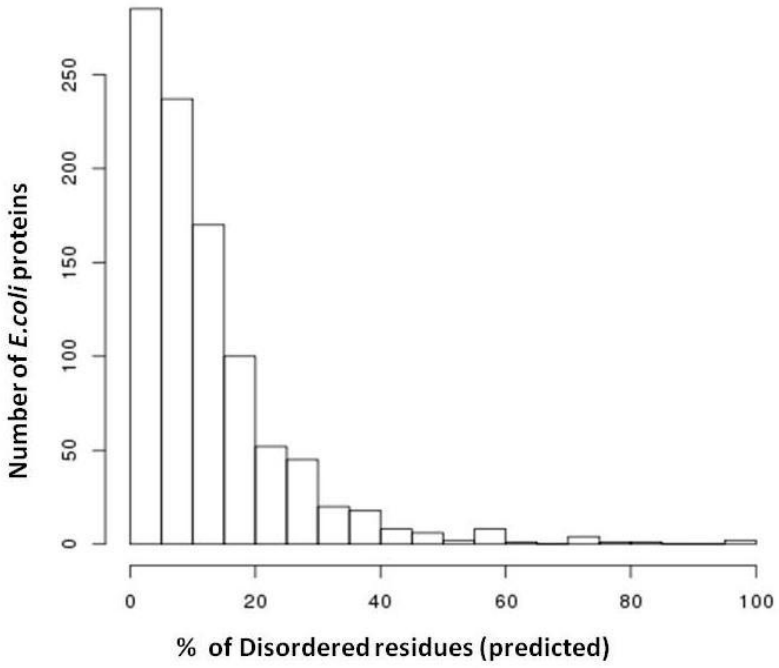


Figure 24. Distribution of predicted percentage structural disordered residues in 1,196 E.coli proteins. Y-axis is the number of E.coli proteins in each bin.

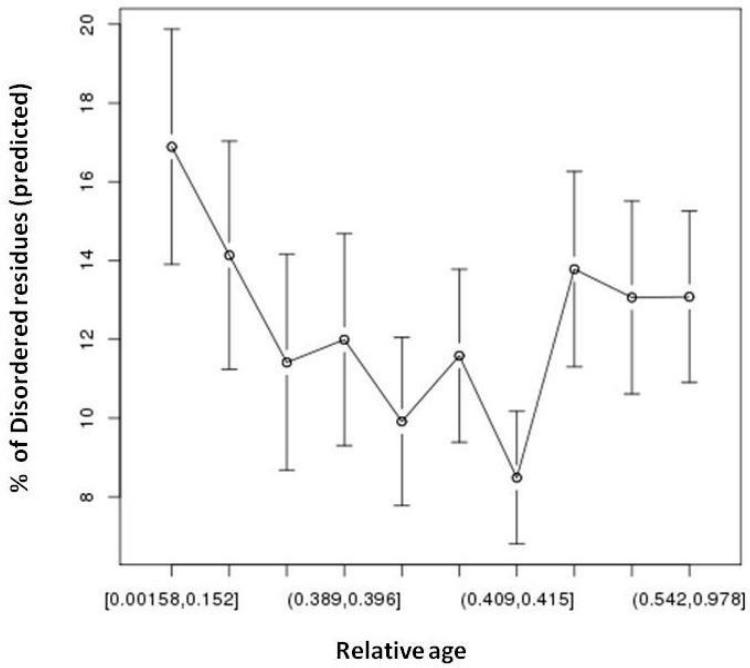


Figure 25. Average predicted % of structurally disordered residues in E.coli K12 proteins as a function of family age.

Figure 25 shows the predicted disorder in these 1,196 E.coli K12 proteins as a function of the relative age of their families. Variances here are high, but the proteins in the youngest families are predicted to have the greatest amount of disorder. Disorder falls off with age, by a total factor of about 2, before rising again for proteins in the oldest families (P-value = 8e-02 by Pearson correlation and P-value = 7 e-02 by Kendall Tau correlation).

Subsection 7. Cross-correlations among the observations

Table 1. Comparison of Pearson correlation analysis (Pearson P value) between two pair of protein properties (x and y) and the corresponding Partial correlation analysis result. For each pair of variables, the Pearson correlation P value is given (between x and y), followed by the P value after controlling for a third variable (z). The next column notes the type of correlation (positive, negative or no correlation). The last two columns note when the control variable was masking the extent of the correlation between the other two variables, or artificially enhancing it. Partial correlation values, indicating that the simple correlation is to some degree an artifact of correlation with a third variable, are in red

Pair of correlation (x and y)	control variable (z)	Pearson P value	Partial Correlation P value	Correlation (+/-)	Masked by	Artifact of
Age -expression		$< 2.2*10^{-16}$		+		
	Disorder		$1.5*10^{-8}$			
	Rate		$9.7*10^{-23}$		Rate	
	PPIs		$7.1*10^{-56}$		PPIs	
Age-Rate		$<2.2*10^{-16}$		-		
	Disorder		$4.9*10^{-19}$			
	Expression		$4.9*10^{-13}$			
	PPIs		$1.2*10^{-19}$			
Age-Disorder		$8*10^{-2}$		none		
	Rate		$9.0*10^{-2}$			
	Expression		$3.4*10^{-5}$		(weak) expression	
	PPIs		$5.0*10^{-3}$			
Age-PPI		$6.45*10^{-7}$		+		
	Disorder		$4.6*10^{-12}$		Disorder	
	Rate		$1.76*10^{-7}$			
	Expression		$9.0*10^{-4}$			expression
Expression-Disorder		$2.03*10^{-10}$		+		

	Age Rate PPIs		$1.0*10^{-8}$ $1.6*10^{-11}$ $2.5*10^{-3}$			PPIs
Expression-Rate	Age Rate PPIs	$2.39*10^{-7}$	0.11 $2.4*10^{-8}$ $1.2*10^{-7}$	-		Age
Expression-PPI	Age Rate Disorder	$3.03*10^{-9}$	$2.4*10^{-7}$ $1.4*10^{-9}$ $2.1*10^{-12}$	+	(weak) disorder	Age
Rate-Disorder	Age Expression PPIs	$1.18*10^{-5}$	0.17 0.02 0.4	(weak)-		Age Expression PPIs
Rate-PPI	Age Expression Disorder	$6.73*10^{-7}$	0.06 0.18 0.87	+		Age Expression Disorder
Disorder-PPI	Age Expression Rate	$1.8*10^{-10}$	$1.6*10^{-9}$ 4.7 $2.7*10^{-6}$	(noisy)+		Expression

The analysis so far has established correlations between four quantities (expression, evolutionary rate, disorder, and number of protein interactions) with apparent family age. However, correlations of these quantities do not prove that in some sense age is the determinant of these effects. In particular it may be that correlations of some of these properties with age are artifacts of age and the property of interest being correlated with a third property. To investigate this possibility, we determined the partial correlation of each pair of variables when the effect of a third is removed[76]. Table 1 shows the results.

As the table shows, in addition to the correlations already discussed, there are a number of others between pairs of variables. For example, expression correlates strongly with evolutionary rate, disorder and protein interactions. Examination of the partial correlations reveals two main points: First, the correlations between age and the other four variables - expression level, evolutionary rate, disorder and protein interactions (PPIs) – are not weakened when controlling for cross-correlations, with the exception of a moderate weakening of age/PPI when controlling for expression. In fact the age/expression correlation becomes substantially stronger when the effects of cross-correlation with rate and PPIs are removed. Secondly, all other simple correlations are substantially weakened when one or more of the other variables is controlled for. Most strikingly, the correlations of rate with disorder, expression, disorder and PPI are all seen to be a complete artifact of cross-correlation with third variable, particularly age. Overall, the results are consistent with 'age' as the driving valuable in the observed set of correlations.

Section 4 Discussion

We have examined the relationship of four properties related to the structural and functional maturity of proteins, as a function of the age of the corresponding protein families. The results are consistent with the hypothesis that by these measures, overall, apparently young proteins are immature, and support a model in which maturation of new proteins is slow on an evolutionary time scale.

We find that average mRNA expression levels in E.coli K12 are strongly dependent on family age, with expression 16 fold larger in the oldest 1/8 of families compared

with the youngest 1/8 of families. This observation is consistent with young proteins being poorly adapted to high concentrations. It also offers an explanation of why young proteins are difficult to work with experimentally[106], since expression is normally done at high concentrations. A previous study in yeast[57] also observed a positive correlation of expression and apparent age.

We also find that the rate of sequence change in the youngest families is on average substantially larger than in the oldest ones, by a factor of five, with the rate in the youngest 1/8 of families more than a factor of two faster than the next youngest set. This observation is consistent with the view that the youngest proteins are under strong positive selection, and still maturing in terms of structure and function. A similar relationship between family age and evolutionary rate has been noted in higher Eukaryotes[53, 56, 65]. Variable rates of sequence change in young proteins have also been interpreted as indicating more variable selective pressures than in older proteins[, 54, 55, 66].

The number of known protein-protein interactions (PPIs) increases substantially with age, from an average of about one for the youngest families up to more than five times as much for the oldest subset. Generally, each new protein-protein interaction represents a new function of the protein, so that the data suggest an ongoing acquisition of function that is likely not yet complete in many cases. Kunin et al. also noted that proteins of different ages have different connectivity levels in interaction networks in yeast[69].

Finally, there is a more complex dependence on the fraction of predicted structural disorder and age. E.coli K 12 proteins belonging to the youngest protein families do exhibit the highest levels of disorder, consistent with incomplete evolution of the tertiary structure. There is a steady decrease with increasing age to the point where the average fraction of disorder is halved. But the oldest families have an intermediate average fraction of disorder. It has been observed that disorder increases with the number of protein interactions[73], and the suggested cause is that as the number of interactions increases, segments of proteins become more disordered to allow interaction with multiple partners. For example, a short disordered segment in Human P53 interacts with four different partners, in each case adopting a different conformation in the complex[69].

While these four measures of protein maturity all correlate with the corresponding family age, correlation does not establish a causal relationship with age. In particular, since there are significant correlations between the four properties as well (Table 1), it is not clear which underlying effects cause which observations. To address this point, we performed a set of partial correlation analyses[74], examining the effect of removing the influence of each factor on correlations between each pair of variables. The correlations of each of the four factors with age are largely unaffected by the removal of the influence of the other variables. (Although the correlation of the number of protein interactions with age is slightly weakened when controlling for expression, it is still strong). Others have noted that rate and age remain correlated when controlling for expression level[54]. Conversely, apparent correlations between rate and expression, rate and disorder, and rate and PPIs are all seen to be an artifact of the correlations of

each of these variables with age. A previous study found a weak positive correlation between predicted disorder and expression level[108]. These results suggest that is an artifact of correlations with age.

Of particular interest is the observation that the negative correlation of expression and evolutionary rate is an artifact of each of those quantities correlating with age. Figure 26 shows the relationship between expression and rate. A negative correlation of expression with evolutionary rate has been found for a number of species[57], and gene expression level has been described as an important constraint on the evolutionary rate of proteins[58]. Several hypotheses have been proposed to explain this observation. Drummond and Wilke have asserted this relationship is a result of selective pressure for translational robustness because levels of mistranslated protein increase as gene expression level increases[61, 62, 63]. While attractive, there is little data to support this explanation. At least for these E.coli proteins, it appears that the correlation between expression and rate is derivative on the correlations of both with age, suggesting that the 'negative selection because of difficult folding' hypothesis may not be correct. Examination of these relationships for other organisms is required to further substantiate this point.

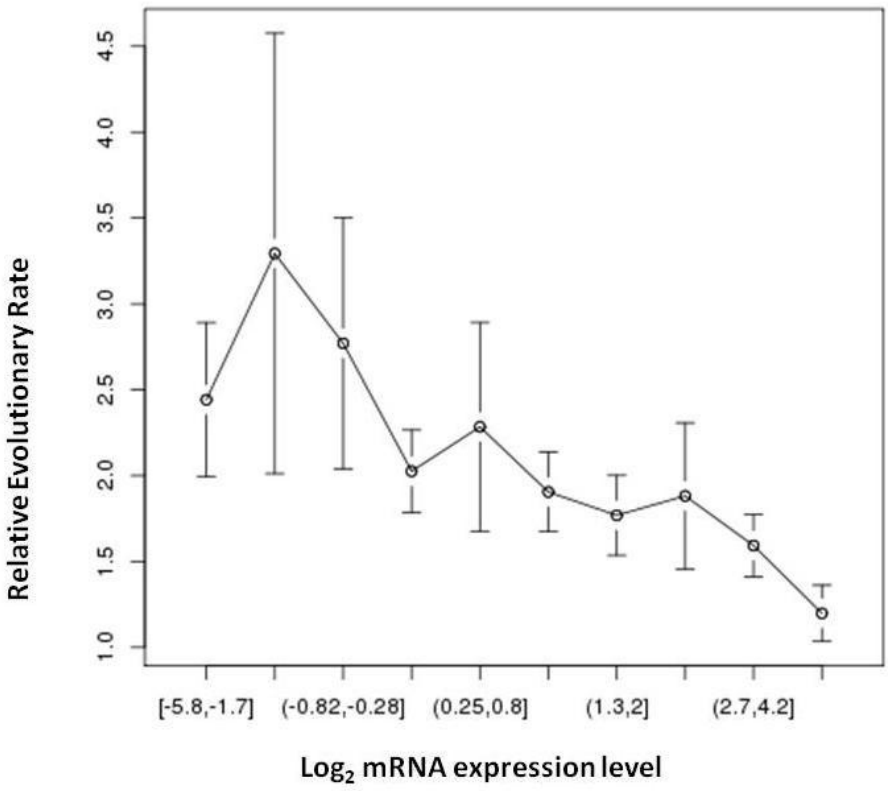


Figure 26. Comparison of relative evolutionary rates of sequence change in orthologous families and mRNA expression level for 514 proteins in E.coli K12. (Expression data are from http://www.genome.wisc.edu/, log phase growth on glucose. Y axis shows relative evolutionary rates in orthologous families, bars show 95% confidence intervals (approximately 2 sigma). X axis has equal points per expression bin).

The analysis was performed on a set of prokaryotic genomes, and as expected[29, 30, 31], the level of lateral gene transfer among these organisms is substantial. We have partially corrected for the influence of this process on apparent family age using a procedure that removes proteins for which the sequence relationship to other family

members is inconsistent with linear evolutionary descent. As figures 19a illustrates, the correction generally leads to a smoother relationship between age and each of the four factors considered. It is clear that the conservative removal process used does not come close to completely eliminating the effects of LGT. Nevertheless, in spite of the remaining noise, reasonably straightforward relationships with age are revealed.

There are other possible causes of artifacts in the data. For example, it is more difficult to experimentally detect protein complexes for less highly expressed proteins, and since expression increases with age, this could account for the apparent increase in PPIs with age. Table 1 shows there is a correlation of expression and PPIs, but that this is an artifact of both correlating with age. As noted earlier, the PPI/age correlation is only slightly weakened by removal of the effect of expression. Both observations support the view that there is a real dependence of PPI level with age.

In summary, for the properties examined, there is strong case for a substantial fraction of proteins in apparently young families indeed being of recent origin, and of a slow maturation of both the functional and structural properties of many proteins. Further studies are desirable to reinforce these findings, particularly including a larger number of genomes, examining the properties of Eukaryotic families, and extending the factors considered to include aspects of tertiary structure, such as fold class and use of local structural motifs.

Section 5 Material and Methods

Subsection 1. Orthologous protein domain families

The analysis was based on a previously compiled set of 30,658 primary orthologous protein domain families (of which 4,856 have three or more members) compiled from all annotated open reading frames in a set of 66 prokaryotic genomes[8]. We use the subset of 1,196 of these families which have a member in E.coli K12. These families include 94 singletons (families with only one member), 151 doubletons (families with two members) and 951 primary orthologous families containing three or more members. All analysis was performed on this orthologous set.

Subsection 2. Calculation of the relative age of each orthologous family

The relative age of each of the 1,196 orthologous protein families was deduced from a previously constructed species tree (Chapter 2). The tree was built using information from a large number of the orthologous domain families, rather than the more common procedure of incorporating data from only a few highly conserved families. As a result, the tree branch lengths are expected to be better determined and exhibit less bias, providing a stronger foundation for the present analysis. Branch lengths are in units of amino acid substitutions per site in a reference set of 14 conserved protein families, thus a branch length of 0.1 is the interval in which an average of 1 substitution per 10 residues will have occurred in those families.

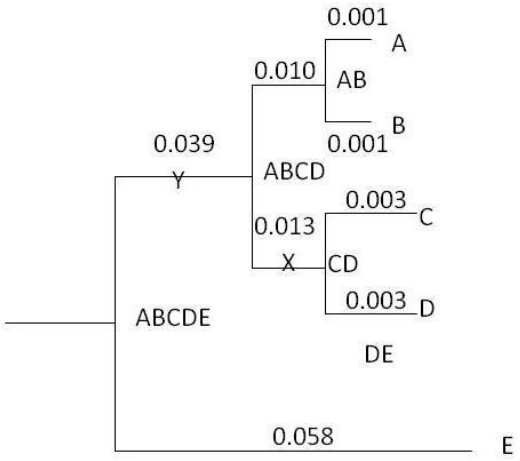


Figure 27 Phylogeny based estimation of protein family age. The most likely origin of each family is assumed to be at the mid-point of branch below the sub-tree that includes all family members. Thus, a family with members only in genomes C and D is considered to have an origin at the point 'X', and the apparent age is the average total branch length from that point the top of the tree (in this case, 0.0095 substitutions per site). A family with members in, say species in A, C and D has its most likely origin at the point Y, and so an apparent age of 0.033). In this example, species A, B, C, D and E are S. typhimurium, S.enterica, E.coli O157, E.coli K12 and Y. pestis respectively, in the current species tree.

The apparent age of each orthologous family was estimated from its phylogenetic distribution, as illustrated in figure 27. Lateral gene transfer can result in an apparent increase in family age. Consider a family that originated somewhere along the ABCDE-to-ABCD branch, with members in species A, B, C and D. The most likely

age is 0.033, the average total branch lengths from Y to the top of termini of branches A, B, C and D.

If there is a gene duplication and transfer to species E, the apparent age will be at the root of the tree, and so greater than 0.058, nearly twice as old. To partially correct for distortion of apparent age by LGT we applied a previously developed method that detects which genes have most likely undergone LGT. The principle of the method is that lateral gene transfer of a gene will result in the sequence differences between it and other members of the same orthologous family being inconsistent with those expected from the species phylogeny[10]. The method is applicable to protein families with five or more members. Proteins with such anomalous rates were removed from each family before the calculation of family age. In all, 4,675 proteins were removed from 528 families, affecting their apparent age.

Subsection 3. E.coli mRNA Expression level data sources

A dataset of $\log_2$ mRNA expression level for *E. coli* K12 genes was retrieved from a set of *E.coli* MG1655 microarray-based gene expression profiles obtained under normal growth conditions (LB medium with 0.4% glucose at 37°C) (http://www.genome.wisc.edu/ Wei Y et al. paper[111]). 971 of the 1,196 families included in the analysis have measured *E.coli* K12 mRNA expression levels, and so are used in the expression related work.

Subsection 4. Estimation of protein families' evolutionary rates

Previously calculated relative evolutionary rates (Chapter 2) are used in the analysis. As discussed in Chapter 2, rates are restricted to those where consistent values were

obtained using three different noise resistant methods, to increase the reliability of the results. 514 families out of the 1,196 primary orthologous families with member from E.coli K12 have reliable rates by this criterion.

Subsection 5. Determination of predicted percentage protein disorder

For each of the included 1,196 E.coli K12 proteins, the percentage of disordered amino acids was estimated using DISOPRED2[73], with the default setting of a 3% false positive rate. DISOPRED2 is trained with experimental observations of disordered and ordered residues in protein crystal structures. The predictor uses protein sequence as the input and returns. DISOPRED2 initially runs a PSI-BLAST search of the query sequence over a filtered sequence database. The position-specific scoring matrix at the final iteration of PSI-BLAST is used to generate inputs to the support vector machine (SVM) classifier, returning an assignment of disordered (positive) or ordered for each residue .The method has been assessed in blind tests and found to have a false positive rate close to that claimed[110].

Subsection 6. Protein- protein interaction dataset

Two datasets of experimentally observed physical protein-protein interactions obtained from a high throughput screen using TAP-TAG technology were downloaded from www.bacteriome.org[112]. These consist of a ‘core’ dataset of 4,863 interactions between 1,100 proteins and an ‘extended’ dataset of 9,860 interactions between 2,131 proteins which includes the ‘core’ set and an additional set protein-

protein interaction deemed to be of slightly lower quality than the core set. The extended set was used for our analysis.

Subsection 7. Statistical analysis

Pearson correlation[113] and Kendall Tau correlation[114] were used to estimate of the strength of the linear dependence between each pair of variables. Partial correlations[76] were used to assess the degree of association of each pair of variables, removing the effect of a third. Statistical analysis was carried out in the R package.

Chapter 4: Composition bias and the origin of ORFan genes

Section 1 Abstract

Motivation**:** Intriguingly, sequence analysis of genomes reveals that a large number of genes are unique to each organism. The origin of these genes, termed ORFans, is not known. Here, we explore the origin of ORFan genes by defining a simple measure called “composition bias”, based on the deviation of the amino acid composition of a given sequence from the average composition of all proteins of a given genome.

Results: For a set of 47 prokaryotic genomes, we show that the amino acid composition bias of real proteins, random "proteins" (created by using the nucleotide frequencies of each genome), and “proteins” translated from intergenic regions are distinct. For ORFans, we observed a correlation between their composition bias and their relative evolutionary age. Recent ORFan proteins have compositions more similar to those of random “proteins”, while the compositions of more ancient ORFan proteins are more similar to those of the set of all proteins of the organism. This observation is consistent with an evolutionary scenario wherein ORFan genes emerged and underwent a large number of random mutations and selection, eventually adapting to the composition preference of their organism over time.

Section 2 Introduction

The work was done with our collaborator, Dr. Ron Unger, from The Minaand Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel and his former student, Inbal Yomtovian at the Department of Computer Sciences, Bar-Ilan University, Ramat-Gan, Israel. The main idea of this work is to investigate the origin of the ORFan genes or singletons in terms of the composition bias properties of proteins as a function of their age. Regarding my contribution to the work in this publication: all phylogenetic tree construction and estimation of the relative age of ORFans (Supplementary figure S3) was done using my phylogenetic tree analysis, an earlier version of that described in Chapter 2, using many proteins families, but only least median squares to determine relative evolutionary rates and the recursive filtering method to determine intergenome distances. My work also included the calculation of the Pearson correlation coefficient between two properties (number of ORFans and percentage of ORFan genes with relative age) as shown in figure 30a and 30b. My third contribution was the correlation between the relative ages of the ORFans with various measures related to composition bias of ORFans with either regular proteins (Supplementary figure S4a) or random proteins (Supplementary figure S4b) in particular species. We also compared the correlation between the relative age and the ratio of the overlap of ORFans with either real proteins or random proteins (Supplementary figure S4c). The result and discussion part of this publication have been done during Dr. Ron Unger sabbatical in our lab during 2009 at Institute for Bioscience and Biotechnology research under University of Maryland, College Park, so that I was involved in those aspects of the work as well.

One of the consistent and intriguing observations that emerged from the extensive availability of whole genome sequences is the large number of genes that seem to encode unique proteins that do not exist in other organisms, or exist only in very closely related organisms. This appears to be the case even when using sophisticated sequence comparison methods like psi-blast. These genes are commonly called ORFan genes[6] and the resulting proteins are called ORFan proteins. It was estimated[14] that 20-30% of the open reading frames in a given genome are ORFans. These observations were made early in the history of genome analysis, when only the first organisms had been sequenced. At that time, the common explanation was that these "unique" genes were not unique at all, but that not enough organisms had been sequenced to follow the evolution of these genes. However, while the fraction of ORFan genes has somewhat decreased as more genomes became available, it also became clear that the phenomenon is not a mere artifact of a small sample size; rather, even with the availability of the complete sequence of close to a thousand genomes, there remain a large number of genes whose evolutionary history is not accounted for.

Several possible explanations were given over the years for this phenomenon (for a review see reference[19, 115]. One explanation is that those sequences are not real genes; rather they may represent open reading frames that are never expressed. However, several studies have shown[118] that these genes are expressed, and some of the resulting proteins have even been subjected to 3D structure analysis (x-ray or NMR)[14]. Another possible explanation is that these genes came from lateral gene transfer (LGT). In order for this explanation to be logically relevant, the transfer should have come from genomes whose sampling is sparse and thus can serve as a reservoir for the unique genes. Viral and phage genomes have been suggested as such

a reservoir[25], although other recent studies[7] have indicated that LGT cannot be the source for most of these genes. Another possibility that has been suggested[19] is that ORFan genes originated from ancestral genes, but because of fast evolutionary rate, these genes have mutated their sequence to such an extent that their ancestors are no longer recognizable. Yet another possibility is that some ORFan genes emerged *de novo* from non-coding regions of each genome without being inherited in the regular evolutionary path, for example by shifting the reading frame, a phenomenon called overprinting (see e.g. in reference[18]) or by mutations that change non-coding regions to open reading frames[19].

It is well known that protein sequences have different amino acids compositions, i.e. not all of the 20 amino acids appear in proteins with the same frequency of 5%. The composition is different for different organisms[75] and has both evolutionary and functional origin and consequences. Furthermore, within genomes, different sequences have different compositions, and we term the deviation of each sequence from the average composition of the organism as *composition bias*. The composition of sequences has been used as one of the main considerations in predicting the sub-cellular localization of proteins[117]. Furthermore, it was observed[118] that proteins of the same fold but with unrelated sequences have similar amino acid composition, and thus it was suggested that amino acid composition can help to predict structural folds.

In an attempt to shed light on the evolutionary history of ORFan proteins, we explored the composition bias of 47 prokaryotic organisms. Using a simple measure, we compared the composition bias of the set of all proteins, of random proteins and of ORFan proteins in each genome. We show that the tendency of ORFan proteins to

behave like the rest of the proteins increases with the evolutionary age of the ORFans, and we discuss the evolutionary implications of this observation.

Section 3 Results

The list of the genomes and the number of proteins and ORFan proteins in each genome is given in Supplementary material table S1. We started by calculating the composition bias of the proteins translated from the coding genes, from random "genes" (based on the nucleotide frequency of the entire genome, from the antisense strands of the coding genes, and from the intergenic regions of the genome. The histograms of the composition biases are shown in figure 28 for six organisms: *E.Coli. Rickettsia conorii, Treponema pallidum, Corynebacterium glutamicum, Aeropyrum pernix* and *Clostridium acetobutylicum*. As the number of sequences in the intergenic sets is 1/3 of those of the other sets (see Section 5 Methods), their histograms were normalized by multiplying each value by 3. The real proteins have smaller composition bias (as is evident from the fact that their histogram is the leftmost) than the composition bias of the random proteins. This is expected since the compositions are compared with the average compositions of the real proteins. Surprisingly, the composition bias of the antisense proteins is greater than that of the random proteins. We also note that for all organisms the composition bias histogram of "proteins" translated from intergenic regions are further shifted to the right.

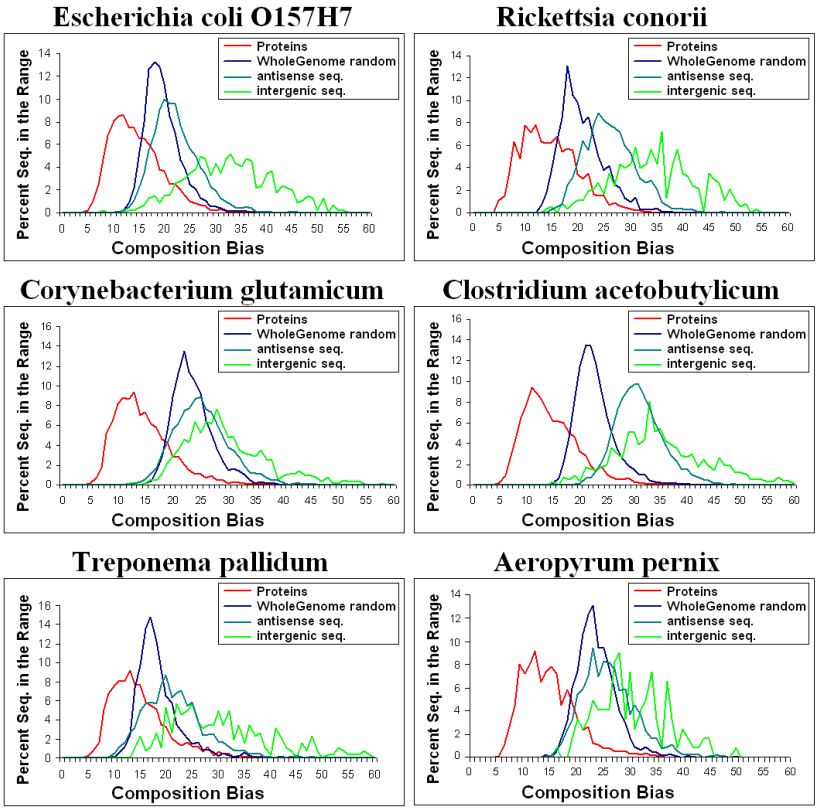


Figure 28. Histograms showing the composition bias for six organisms of several sets of proteins. All histograms were computed by using the average composition vector of the real proteins as the reference, and the composition bias of each protein relative to that reference was calculated. As expected, the real proteins have the smallest bias. Surprisingly, the composition bias of intergenic "proteins" is significantly larger than that of random or anti-sense proteins. For the random genes, very similar results were obtained when using either the genome's coding or non-coding frequencies. (This work has been done by Inbal Yomtovian and Dr. Ron Unger.)

We next compared the composition bias of ORFan proteins to that of the other datasets. Figure 29 shows the composition bias histograms of real proteins, random proteins and ORFan proteins (scaled up to the size of the other groups) for several

genomes. We noticed that ORFan proteins from different species behave differently in their similarity to either the coding or the random groups. The ORFans of E.Coli and Rickettsia conorii look like random proteins (figure 29a) the ORFans of Treponema pallidum and Aeropyrum pernix resemble real proteins (figure 29c) while the ORFan proteins of Corynebacterium glutamicum and Clostridium acetobutylicum have intermediate assignments (figure 29b).

From the results of the calculations for all 47 organisms, we noticed that indeed there is a range in the similarity of the composition bias between the ORFan proteins and the real and random proteins. In an effort to understand this range, we looked at the relative age of the ORFans, as determined by the phylogenetic tree (see Section 5 Methods), as a possible explanation.

First, we checked the correlation between the number of ORFans in each genome and their relative age, and found a weak correlation (0.36). A more significant correlation (0.5) was found between the relative age of the ORFans and the percentage of ORFan genes from the total number of coding genes in the organism (see scatter plots in figure 30).

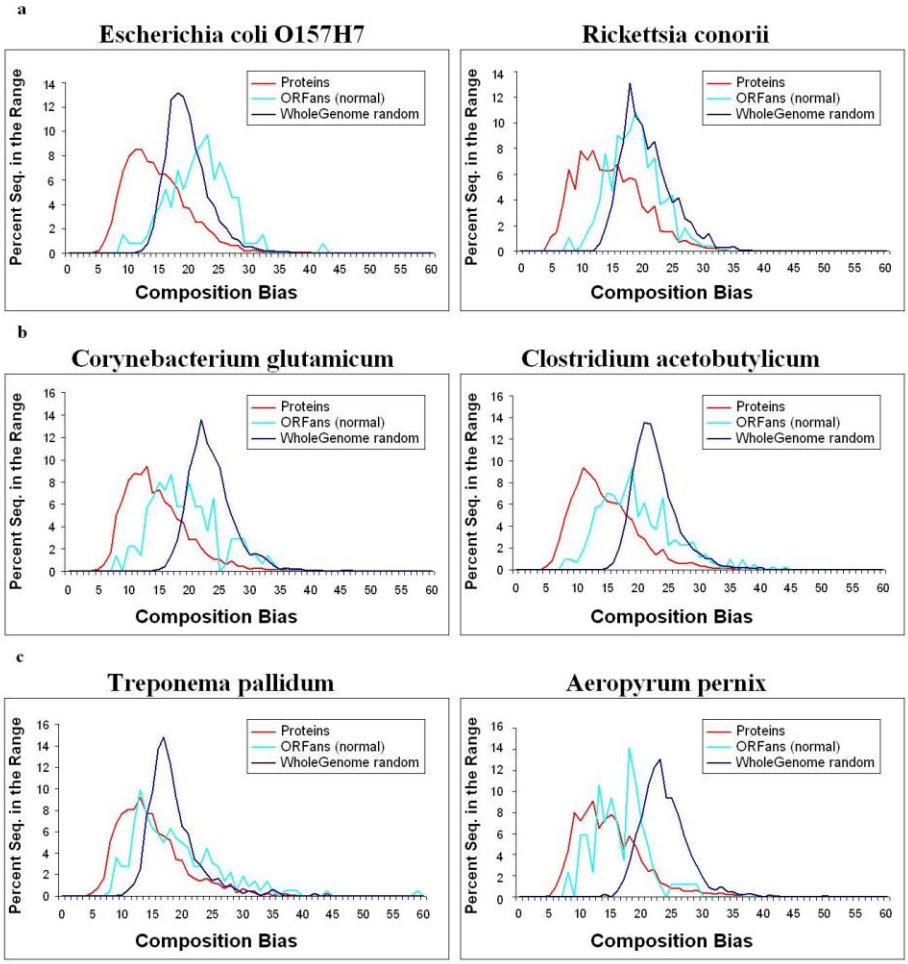
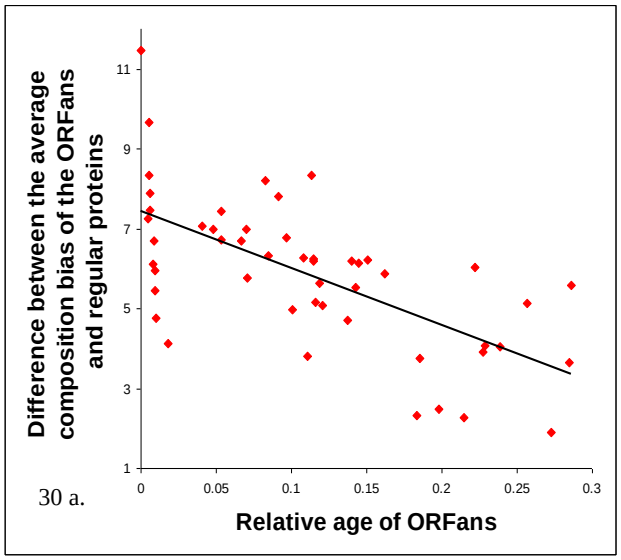


Figure 29. Histograms of the composition bias of the set of ORFan proteins are compared with the composition bias of all proteins and of random proteins for six organisms. Since there are fewer ORFan proteins, their histograms were scaled up accordingly (the results were validated to ensure that they are not due to sampling effects). In the two examples in the top panel (a), the ORFan proteins behave like random proteins; in the two examples in the bottom panel (c), the ORFans behave like the real proteins; and the behavior of the examples in the middle panel (b) is intermediate. (This work has been done by Inbal Yomtovian and Dr. Ron Unger)

Next, we found a surprising strong correlation coefficient of 0.59 between the relative age of the ORFans and the distance between the average composition bias of the ORFan and the random proteins. Similarly, the correlation coefficient between the relative age and the distance between the average composition bias of the ORFan and the real proteins is −0.66 (see scatter plots in Supplementary figures S4a and b). To make sure that these high correlations are not dependent on the particular way of comparing the composition bias, we also calculated the correlation between the relative age and the ratio of the overlaps (and got similar results (correlation coefficient of −0.58, see Supplementary figure S4c).



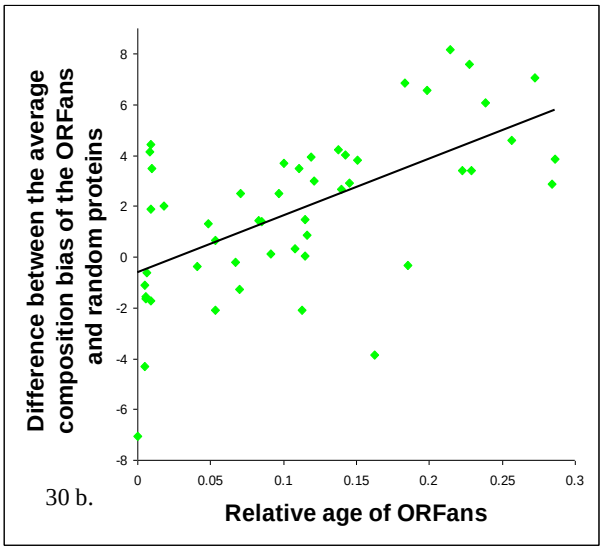


Figure 30. We observed (30a) a weak trend of correlation between the relative age of ORFans and the number of ORFan of the organisms (correlation coefficient of 0.36). A stronger correlation (0.5) was found when the relative age of the ORFan was plotted not against the absolute number of ORFan genes but against the percentage of ORFan genes from the total number of genes of the organism (30b). (Relative age of ORFans in this result was from my phylogenetic tree work analysis)

Section 4 Discussion

The main finding of this study is the correlation between the relative age of the ORFans and the degree of similarity of their composition to that of the real proteins of the organism. We found a significant correlation (correlation coefficients between 0.58 and 0.66) between the relative age of the ORFans and their composition bias, as determined by various measures of the composition distance between the set of the ORFan proteins and the set of real proteins. Thus, the older the ORFans, i.e., the more ancient the organism, the more the amino acid composition of its ORFans resembles that of the rest of the proteins. Young organisms, i.e. organisms that split from their ancestor organisms more recently, tend to have ORFan genes with composition that is more different from that of the rest of the proteins, and more similar to that of the random genes.

We tested to see if there are other factors that correlate with the relative age of the ORFan proteins and with the composition bias. As expected, we found that the fraction of ORFan genes among all coding genes in each organism is correlated with the evolutionary age of the organism (correlation coefficient of 0.5). Older organisms that have, almost by definition, fewer close relatives, tend to have more ORFan genes. No other factors that we tested, including the GC content of the organism, the size of the genome and the ratio of coding to intergenic regions, showed a strong correlation (< 0.3) with the ORFan behavior.

Thus, our data are consistent with a model wherein ORFan genes emerged with a composition that was similar to the random composition of the genome. Then, during evolution and due to the selective pressures that shape the composition bias of each

organism, the composition of ORFan genes gradually converged to be more similar to the composition of the rest of the proteins of the genome.

We may examine the three possible explanations for the origin of ORFan genes in light of this observation. The first explanation is that ORFan genes originated from bacteriophages (see a review in Daubin and Ochman paper[115]). We think that this is unlikely. First, note that bacterial genes that have known homologues in bacteriophage are not considered ORFans by our definition. Second, for six bacteria for which sufficient bacteriophages have been sequenced, we compared the composition of the ORFan genes with the composition of bacteriophage proteins and found that the composition of the ORFan genes of the bacteria is not similar to the composition of the bacteriophage proteins (see Supplementary figure S5).

The second possible explanation is that ORFan genes emerged *de novo* from non-coding regions of the genome (see a review in reference[19]). This is also not consistent with our observation that protein created from intergenic sequences are distinct (further to the right in figure. 28) from the random proteins, while the ORFan proteins fall between the random and the real proteins. If ORFan proteins emerged from intergenic regions, then we would expect the ORFan genes to behave more closely to intergenic non-coding regions of the genome, and not like random sequences.

The third explanation is that ORFan genes result from a very fast evolutionary clock rate of mutations operating on genes that are under positive selection[19]. This explanation is the most consistent with our observations. Random mutations are likely to create nucleotide sequences that have A/C/G/T frequencies that are similar to random sequences, thus creating novel proteins whose amino acid sequences have

composition bias similar to the random proteins that we have created. Over time, the sequences underwent further mutations and selection that changed their composition and brought their composition bias to be more similar to that of the rest of the proteins.

Section 5 Material and Methods

Subsection 1: Dataset

Our dataset started with a collection of 66 representative prokaryotic genomes[10]. For these genomes, the sequences and annotations were taken from NCBI. In each organism, ORFan genes were defined as genes that appear only in their genome-of-origin, and do not have any similar genes based on a Blast run against the entire NCBI-NR database. The parameters used to define a hit were *E*-value < 0.05, and match-length that covers at least 50% of the ORFan length. Three organisms were found to have another related organism with which they share many proteins (*Escherichia coli* with *Shigella*, *Methanococcus* and *Nostoc* sp PCC 7120 with *Anabaena*). For these organisms, we considered genes as ORFans if they appeared only in their original genome and in the very close relative.

The analysis presented here included the 47 genomes (out of the 66) that have at least 25 ORFan genes each. The list includes 38 bacteria and 9 archaea (see Supplementary Table S1). All together, we identified 8812 ORFan genes out of 123 444 genes (~7%) in our ensemble (Supplementary table S1).

Subsection 2: Real and random proteins

We called the set of all proteins in an organism the set of 'real proteins'. For each organism, three sets of random sequences were created. Each set was matched to the set of real proteins in terms of the number of proteins and the length of each protein. The three sets of random sequences were created based on the nucleotide frequency (i.e. the A/C/G/T ratios) of (i) the entire genome, (ii) of only the coding regions and (iii) only of the non-coding regions. The nucleotide sequences were translated to amino acid sequences. All sequences started with ATG, and to maintain protein length, stop codons, when generated, were replaced by other random codons.

Subsection 3: Translating proteins from intergenic regions

Nucleotide sequences that came from intergenic regions of the genome (i.e. regions that are between genes and do not reside on the opposite strand of coding regions) were translated into proteins. Stop codons were skipped over and the subsequent nucleotides were used to create additional codons such that the lengths of these 'proteins' match those of the real proteins. Since the number of intergenic regions in prokaryotic genomes is limited, the set sampled was 1/3 the number of proteins in each genome.

Subsection 4: Translating anti-sense proteins

For each protein, the antisense sequence (i.e. its reverse complement sequence) was also translated. Thus, the size of this set of proteins was the same as that of the real proteins in each genome. Stop codons were skipped over.

Subsection 5: Calculating Composition Bias

For each organism, a reference composition vector was calculated by averaging the percentage of each of the 20 amino acids in each protein over all real proteins of the genome according to NCBI annotation. For each amino acid, the SD about the average composition was also determined. For each amino acid sequence *s*, the composition bias c^s was calculated by comparing its composition vector to the reference composition vector according to:

$$(1) \quad c^s = \sum_i \frac{| f_i^s - f_i^r |}{SD_i^r}$$

Where *i* ranges over the 20 amino acids, f_i^s is the ith component of the composition vector of the given sequence, f_i^r is the ith component of the reference composition vector, and SD_i^r is the standard deviation of the reference composition of the i*th* amino acid about its average. Thus, each "protein" is assigned a composition bias, and for a set of "proteins" in a given organism, we created a histogram of these composition biases. For the ORFan proteins, the histogram was scaled up by a factor based on the fraction of ORFan proteins. For example if an organism has 4000 proteins of which 400 are ORFans, then the values in the ORFan histogram were scaled up by a factor of 10 (4000/400).

We have also compared the frequency vector of the given sequence to that of the reference vector using a root mean square (RMS) measure. The RMS measure square the difference in the frequency of corresponding amino acids without normalization to the SD weight that appear in Equation (1). The results of using these two measures were similar and thus in this article we show only the results of the first measure.

Subsection 6: Calculating the difference between histograms of composition biases

The difference between the histograms was calculated as the difference between the average values of each histogram. We also measured the difference by computing the overlap between the two histograms. We then calculated the ratio between the overlap of the ORFans and real protein and the overlap of the ORFans and the random proteins. This ratio reflects the relatedness between the ORFan proteins to either the real proteins (low ratio values) or the random proteins (high ratio values).

Subsection 7: Phylogenetic tree construction and measuring the relative age of ORFans (This result was from my phylogenetic tree work analysis)

Since ORFan genes are found in only a single branch of the phylogenetic tree, they must have emerged subsequent to the split of that branch. The maximum age of the ORFan genes must be smaller than the age of the organism, and thus it assumed to be proportional to the relative length of their terminal branch (Supplementary Figure S3). This length was used to estimate the approximate relative age of the ORFan.

The tree was constructed incorporating information from accepted amino acid substitutions per site between species in a large set of protein families, to avoid bias issues encountered in methods where only a small number of families is used. The set of orthologous protein domain families previously constructed[8] from 66 prokaryotic genomes was used. Multiple sequence alignments for each family were generated using MUSCLE[94]. The estimated accepted amino acid substitutions per site between each pair of domains '*i*' and '*j*' in each family '*u*', $S(i, j, u)$ were then obtained using

the PROTDIST module in PHYLIP[95] with the Jones–Taylor–Thornton amino acid substitution matrix[96].

The numbers of accepted substitutions per site for each family were placed on the same scale by comparison with the average rates of substitution $S_{ref}(i, j)$ between genomes '*i*' and '*j*' in a set of 14 highly conserved families. The rate of sequence change for each family, $C(u)$, relative the reference set was obtained using a robust least median square procedure[97, 98], finding the $C(u)$ which minimizes the median value of the set $(r(i, j, u)^2)$, where

$$r(i,j,u)^2 = \{S(i,j,u)/C(u) - S_{ref}(i,j)\}^2$$

and the set includes contributions from all pairs of genomes '*i*' and '*j*' with members in family '*u*' [38]. A robust method was necessary to avoid distortions of $C(u)$ arising from anomalous $S(i, j, u)$ values caused by LGT and other factors.

The intergenome distance, $D(i, j)$, between each pair of genomes '*i*' and '*j*' was estimated using $D(i, j) =< S(i, j, u)/C(u)>_u$ where the average includes contributions from all families with members in genomes '*i*' and '*j*'. A phylogenetic tree was then built from this distance matrix, using the neighbor joining method [88], as implemented in PHYLIP.

Correlations were calculated using the standard Pearson correlation coefficient comparing the two properties of interest (e.g. number of ORFans and relative age) for each of the 47 genomes.

Section 6 Acknowledgement

I acknowledge all of the funding: This research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. Partial support from NIH R01GM081511 to J.M., Fulbright Fellowship for Burapha University, University Staff Development Program to N.T. and Israel Science Foundation 1339/08 to R.U.

Chapter 5: Conclusions and Future Perspectives

Section 1 Overview

This thesis reports results of studies of some evolution mechanisms for the possible origin of apparently young protein families found in 66 Prokaryotic organisms. Four possible hypotheses were proposed to account for the emergence of these families: They may arise from (1) previously non-coding DNA, or by a frame-shift in an existing coding sequence; (2) recombination of structural fragments between proteins or recombination with non-coding DNA; (3) older families where the rapid rate of sequence change make relatives hard to detect; and (4) as a consequence of lateral gene transfer (LGT) from other organisms.

The work focuses on obtaining and assessing data relevant to hypothesis (1): that these young families are in some sense new open reading frames, occurring in previously non-coding DNA or as a result of a frame-shift within an existing open reading frame. The basis of the approach is that proteins that have recently arisen in this way will have properties that distinguish them from more established proteins. Thus, we examined five relevant properties as a function of apparent family age. A necessary prerequisite for the analysis is a means of estimating protein family age. The absences of a fossil record as well as the high prevalence of lateral gene transfer make age determination for prokaryotic entities challenging. From previous work, we had established that for our purposes existing phylogenetic methods had significant

deficiencies, as a consequence of only utilizing information from a small number of protein families. We therefore undertook to develop a new method for the construction of species tree that makes use of information from a large number of protein families. The new approach required two major hurdles to be surmounted. First, integrating information from many families required a means of normalizing across the widely varying rates of sequence change among families. Second, the effects of lateral gene transfer and inevitable errors in assigning proteins to orthologous families as well as in sequence alignment introduce large amounts of noise into the data. To overcome these problems, we introduced a combination of noise resistant methods to calculate relative evolutionary rates for each family, and to determine inter-genome distances. With this scope of work, the conclusions presented below fall into five parts: (a) Effectiveness of noise resistant methods; (b) determination and analysis of relative evolutionary rates; (c) development and application of a multifamily phylogenetic method; (d) analysis of protein properties as a function of age; and (e) future prospects.

Section 2 Use of noise resistant methods.

Relative evolutionary rates of sequence change for each orthologous family were based on the ratio of the number of accepted substitutions per amino acid between a pair of family members in two of the genomes included to the corresponding number of average number of amino acid changes between members of 14 highly conserved families. Thus, for a particular family, each pair of genomes that contain members of the family provides one estimate of the rate. Three primary factors introduce noise.

First, if one of the members of a family has recently undergone lateral gene transfer, it will have sequence differences to members in other genomes largely reflecting its original phylogenetic location, not the current one. Second, the method assumes that all the family members included are orthologs. Paralogous protein pairs will typically have larger sequence differences than orthologs in the same species because of adaption to different functions[8, 119]. Many methods have been developed to identify orthologous subfamilies, but none are prefect[8, 120, 121, 122]. Third, obtaining an accurate count of amino acid differences between a pair of proteins requires a correct amino acid sequence alignment. Although alignment methods have improved dramatically in recent years[8, 123, 124, 125, 126], there are still inevitable errors at low sequence identities. Therefore, a straight average of relative rates over all the contributing genome pairs is likely to be seriously distorted by errors of one sort or another. A least squares fit of a rate to the data is also problematic, because of sensitivity of that method to outliers[100]. We used three methods that are more noise resistant to obtain the evolutionary rates. First, least median squares, which is less sensitive to outliers than least squares[97, 98]. Second, a Gaussian Kernel estimator[99]. Gaussian Kernel estimator methods represent each observation with a Gaussian probability density centered at that value, with a variance related to the fuzziness of the data. The sum of Gaussian probability densities for all observations then provides an overall likelihood distribution for the data, and the maximum value is the maximum likelihood estimate. The third method is a recursive filtering procedure, in which a simple average ratio is first calculated. The individual values most different from that average are then removed from the data, and a new average calculated. The process is repeated until the ratio converges

to a constant value, in practice not more than three iterations with these data. None of these methods can deal with the worst cases of noise in these data, most likely arising for families where many lateral gene transfer events have occurred. But for such cases, the errors distort the calculated rates differently for the different methods. We exploited this principle, and selected just those families where the three methods provided closely similar results. 1,379 families out of the 2,264 with more than three members met the consistency criteria used. Least median squares and the Gaussian kernel estimator method agreed for a substantially higher fraction of families than this, with the recursive filtering approach usually being the outlier. The deficiency of that approach is that if the initial average is too much distorted by the noise, inappropriate contributors will be filtered out in the next iteration, and so the initial error is locked in. This particularly tended to be the case for families with few members. The average evolutionary rates were used as weights in combining the data from many families in order to obtain intergenome distances. The same three noise resistant methods and consistency criteria were used in these calculations. The primary source of noise here is errors in the evolutionary rates, and since only the most reliable are included, consistent results should be found in most cases. Reassuringly, this was the case, with 96.9% of the distances meeting the consistency criteria. Based on experience in this work, we conclude that least median squares and the Gaussian kernel estimator are very robust to large amounts of noise in the data and further, the use of a consistency test is effective for identifying those cases where one or both methods are overwhelmed by noise.

Section 3 Determination and analysis of relative evolutionary rates

As others have also observed[53, 55, 58, 59], we found a wide variation in evolutionary rates among the families – well over an order of magnitude difference between the highest and the lowest rates. At present, factors determining the rate of sequence change of different families are relatively poorly understood. An important conclusion from our analysis of correlations of various factors with rate of sequence change (Chapter 3) is that, at least for these data and contrary to what others have asserted[55, 58, 59], the level of gene expression is not a significant controller of evolutionary rate. Instead, it appears that on average, the younger the protein family, the more rapid the rate of sequence change. There are two possible explanations for this. One is simply that there are few functional constraints on young proteins because they have no or weak function. While that could be the case for a small fraction, we have seen (figure 18) that most are significantly expressed, implying it is unlikely that they have no function. More probably, function is still emerging or undergoing fine-tuning, leading to positive selection, resulting in a higher rate of acceptance of substitutions.

We also found one unexpected factor related to evolutionary rate: E.coli K12 proteins in 17 families out of 58 families with rates higher than 5.0 have increased expression (1.05-6.68 relative $\log_2$ mRNA expression level) under stress conditions (cold (16°C), heat (50°C), oxidative stress and glucose-lactose shift) compared to normal growth conditions (glucose)[127, 128]. In contrast to that, only 60 such families with E.coli K12 members out of the 456 in the lower rate group (rate less than 5.0)

have a reported increase in expression under stress conditions, a more than two fold lower level, and significantly different by a chi-squared test (P-value < 0.0012). Some examples of these high rate E.coli proteins are as follows: trimethylamine N-oxide reductase also called energy metabolism cytochrome C type protein, chitoporin protein, which regulates the uptake of chitosugar and Ner-like regulatory protein expression. Further details are given in Supplementary Figure S2. The reasons for the tendency for stress regulated proteins to have higher rates of sequence change are not clear at this point, and the phenomenon requires further investigation.

Section 4 Development and application of multifamily phylogenetic methods

Many methodological advances for sequence based reconstruction of phylogenetic relationships have been introduced[129]. The most popular methods at present use maximum likelihood approaches in which each amino acid position in a multiple sequence alignment is treated as a feature included in the optimization. While powerful, these methods are very computationally demanding, and so are unable to deal with alignments for a large number of families. Hence, in this work we used older methods, first calculating intergenome distances based on average differences between sequences, and then using standard neighbor joining tree-building methods[88]. The result is successful in the sense that the new tree corrected the two primary defects we are aware of in the previous trees constructed using a small number of families – short branch lengths are better resolved because of the inclusion of fast evolving families, and distortion of the relationship between bacteria and archaea is reduced through reduced reliance on ribosomal proteins. The topology of the new tree

is generally robust as assessed by bootstrapping, demonstrating the new method produces stable results. The topology obtained is similar to that of the earlier trees, with a few interesting differences. In particular, the three included hyperthermophiles, which previously clustered together, become separated. We hypothesize that this is in fact a better representation of the true relationship between the species concerned, and that earlier results were artifacts resulting from temperature based amino acid composition bias. This idea requires further investigation.

Section 5 Analysis of protein properties as a function of age

All five of the protein properties examined show significant variation as a function of apparent family age. The extent of correlation was enhanced by use of a technique developed by a previous student[38] to remove the more easily detectable instances of lateral gene transfer that distorted family age. While this treatment is only partial, we have clearly demonstrated its value in phylogenetic age analysis. The principal conclusions from the property/age analysis are as follows: (a) There is strong (16 fold) increase in average expression level as function of family age, consistent with young proteins being as yet poorly adapted. (b) As discussed above, the dependence of average family evolutionary rate on family age is consistent with positive selection for still emerging structural and functional properties, as would be expected for new open reading frames. (c) The average number of known protein binding partners also increases with age, consistent with newly formed young proteins having limited function, and a gradual increase in functional complexity with age. (d) The picture for intrinsic structural disorder is more complex, but is consistent with young proteins

having poorly ordered tertiary structure that gradually becomes better defined with age. Then, as older proteins acquire multiple binding partners this process is partially reversed[70, 72]. (e) Codon usage of the youngest orphans is close to that of random proteins (Chapter 4), as would be expected for new open reading frames. Notoriously, correlation does not imply causation. However, the partial correlation analysis we performed does suggest that age is the driving variable for the behavior of these properties.

While these five results are consistent with the new open reading frames hypothesis, they do not provide conclusive proof. In particular, it is necessary to examine to what extent the observations provide evidence against the other hypotheses: Hypothesis (2), that the apparently young proteins are the result of recombination of parts of older proteins and partially from recombination of these with non-coding DNA could produce some tendency for some of the observed dependencies on age – newly combined structure might exhibit more disorder, initially have higher rates of sequence change as a result of positive selection for new structural features, and also be insufficiently adapted for high expression levels. The very strong resemblance of codon use to that of random proteins would not be expected, though. Also, many previously existing protein interactions would likely be mostly preserved in the recombination process, inconsistent with observed strong dependence of number of binding partners on apparent age. Hypothesis (3), that these are old proteins that have fast rates of sequence change so more distance relatives cannot be detected is fully consistent with the observed dependence of rate on apparent age, but not with any of

the other four factors. Hypothesis (4), that these young proteins are largely the result of lateral gene transfer, is inconsistent with all five observations. That said, it is clear that some fraction of these apparently young proteins are the result of lateral gene transfer from phage – using a PSI-BLAST search against phage sequences in the NR database[130], we find 7.2% of singletons to have a detectable relationship to phage. Indeed, it is likely that all four hypotheses have some validity, but that new open reading frames play the major role.

Section 6 future prospects

Subsection 1. Phylogenetic analysis

The new tools for the utilization of large numbers of families and removal of lateral gene transfer effects in building phylogenetic trees open the way for several interesting studies that were previously difficult, if not impossible. First, the robustness conferred by many families should make it possible to look at the distribution of evolutionary rates for all families in each branch of a species tree. The resulting insight into which proteins changed most as each speciation event occurred should shed new light on the adaptations involved. Related to that, it should also be possible to examine how rapidly each family changed in different parts of the species tree – where did substantial adaptation to new conditions or function occur within each family. A third possibility is to build trees based on the families involved in a particular pathway or GO[93] process: is it now possible to see where these higher level functional entities underwent most change? We also expect that the combined noise

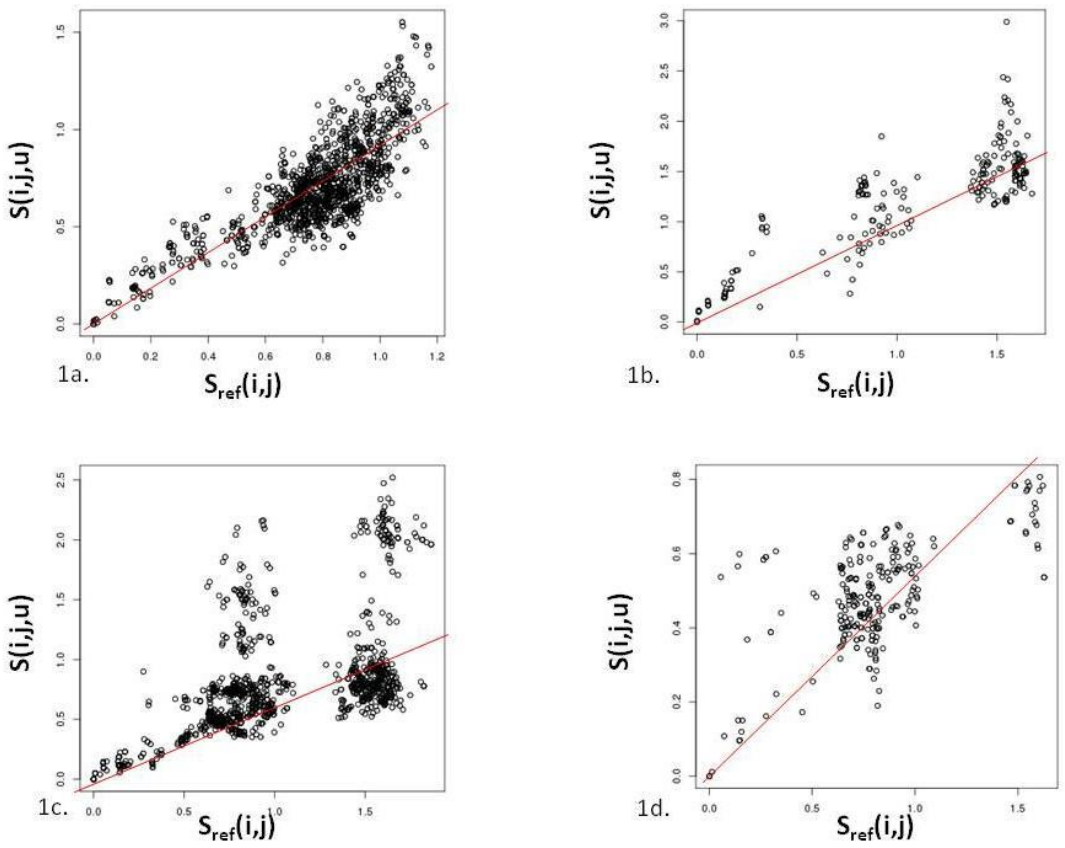
resistant methods strategy developed in this work will find wider application in bioinformatics.

Subsection 2. Further studies of apparently young proteins.

There is still much to be done in investigating the origins of apparently young proteins. First, the new methods are computationally efficient, and can easily be extended to a much larger set of prokaryotic genomes, providing a better quality species tree than those now available. Second, it will be informative to apply the new methods to age dependency for Eukaryotic proteins. Do the same trends as a function of age hold? If so, this would lend further support to the new reading frames hypothesis. Second, new reading frames must come from some where, and an aggressive focus on mapping to previously non-coding or frame-shifted origins is an obvious next step that if successful will provide very strong support for the new frame hypothesis. Third, an additional factor that should be related to protein age is protein structure, so far only examined in terms of intrinsic disorder. We expect that new proteins will exhibit additional structural immaturity, especially in the use of energetically sub-optimal and therefore relatively rare local structural motifs. The difficulty here is that it has proven very difficult to experimentally determine the structures for young proteins[106]. An alternative approach is provided by fragment modeling methods[131], which should be directly applicable to this problem. Finally, if these proteins are new, what type of functional roles do they play? In many cases, of course, function is not known, but as more data accumulate our new ability to more

reliably assign family age will provide a better means of tracking the functional classes of young proteins.

Appendices: Supplementary tables



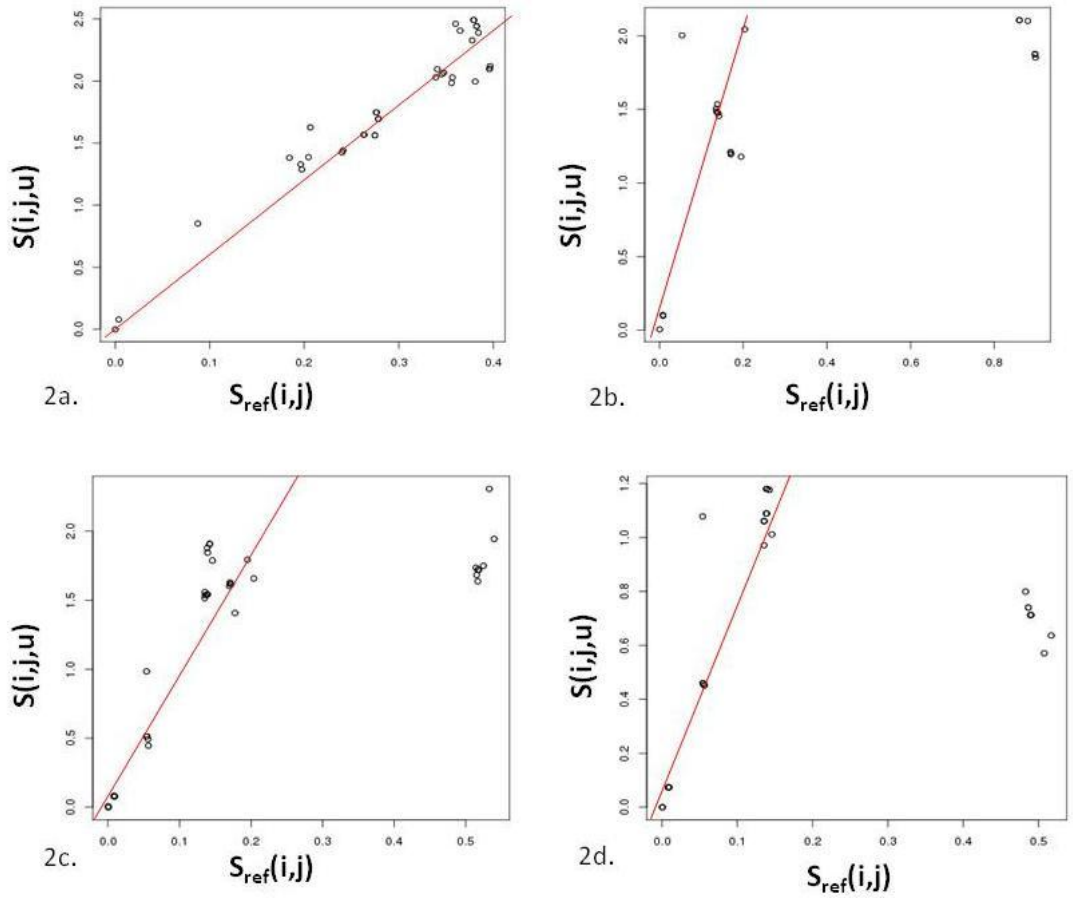
Supplementary Figure S1. Examples of determining the relative evolutionary rate for four protein families using least median squares (LMS), for cases where the rate is less than 1. The LMS line is shown in red, and the slope gives the relative rate, $R_{LMS}(u)$.

1a. Family of Ribosomal proteins L7/L12, with an average relative evolutionary rate of 0.86 (LMS 0.83, GDKE 0.86 and RF 0.89) and 54 members in the family.

1b. Family of translation initiation factor SUI1, with an average relative evolutionary rate of 0.95 (LMS 0.95, GDKE 0.93 and RF 0.98) and 21 members in the family.

1c. Family of Succinyl-CoA synthase, alpha subunit, with an average relative evolutionary rate of 0.61 (LMS 0.59, GDKE 0.59 and RF 0.64) and 49 members in the family.

1d. Family of Uriease (gamma subunit), with an average relative evolutionary rate of 0.58 (LMS 0.57, GDKE 0.59 and RF 0.59) and 23 members in the family.



Supplementary Figure S2. Examples of determining the relative evolutionary rate for four protein families using least median squares (LMS), for cases where the relative rate is greater than 5. The LMS line is shown in red, and the slope gives the relative rate, $R_{LMS}(u)$.

2a. Family of Cell division initiation protein, require for vegative and sporulation septum formation, with an average relative evolutionary rate of 6.15 (LMS 6.21, GDKE 6.07 and RF 6.17) and 9 members in the family. The $\log_2$ of mRNA expression level of this protein in B. Subtilis under normal growth conditions (glucose) is -1.22 while under stress conditions of acid, heat, salt, and cell envelope stresses, the expression level is increases by 1.8, 1.5, 2.7 ,and 2.8 folds respectively (Eiamphungporn W. and Helmann J.D. 2008).

2b. Family of Trimethylamin N-oxide reductase, energy metabolism, cytochrome C type protein, with an average relative evolutionary rate of 9.75 (LMS 9.66, GDKE 9.97 and RF 9.64) and 7 members in the family. The $\log_2$ of mRNA expression level of this protein in E.coli K12 under normal growth conditions (glucose) is -2.21 while under stress conditions of cold and oxidative stress, the expression level increases 0.98 and 1.05 fold respectively (Kang Y. et al. 2005; Jozefczuk S. et al. 2010).

2c. Family of Outer membrane protein slp precursor, with an average relative evolutionary rate of 9.82 (LMS 9.79, GDKE 9.37 and RF 10.30), and 9 members in the family. The $\log_2$ of mRNA expression level of this protein in E.coli K12 under normal growth conditions is -0.66 while under stress conditions of cold and heat stress, the expression level increases 3.24 and 3.54 fold respectively (Kang Y. et al. 2005; Jozefczuk S. et al. 2010).

2d. Family of Ner-like regulatory protein and Ner repressor protein of phage Mu, with an average relative evolutionary rate of 8.07 (LMS 8.07, GDKE 8.05 and RF 8.08), with 8 members in the family. The $\log_2$ of mRNA expression level of this protein in E.coli K12 under normal growth conditions is -1.22 while under stress

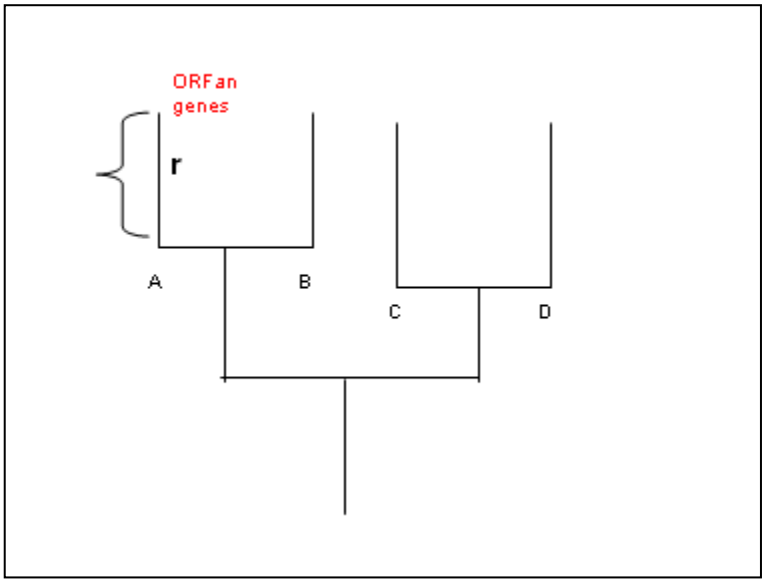
conditions of heat and cold stress, the expression level increases 1.60 and 2.04 fold respectively (Kang, Y. et al. 2005; Jozefczuk, S. et al. 2010).

Supplementary figure S1 and S2 References

Eiamphungporn W. and Helmann J.D. (2008).”The Bacillus subtilis σM Regulon and its Contribution to Cell Envelope Stress Responses.” Mol Microbiol.; 67(4), 830–848.

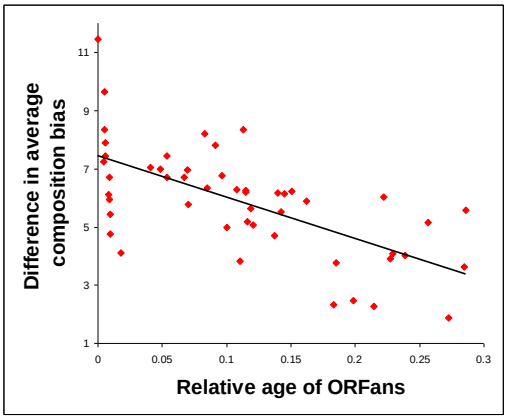
Jozefczuk, S., Klie, S., Catchpole, G., Szymanski, J., Cuadros, I.A., Steinhauser, D., Selbig, J. and Willmitzer, L. (2010). “ Metabolomic and transcriptomic stress response of Escherchia coli.” 6(364), 1-16.

Kang Y., Weber K.D., Qiu Y.,Kiley P.J., and Blattner F. R. (2005) “Genome-wide expression analysis indicates that FNR of *Escherichia coli* K-12 regulates a large number of genes of unknown function.” J Bacteriol 187(3), 1135-1160.

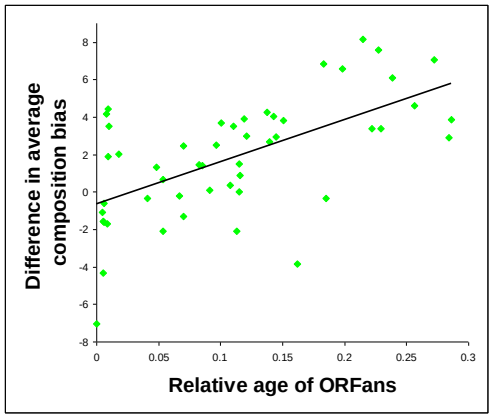


Supplementary Figure S3. Relative age of ORFan proteins. ORFan genes found only in organism A are assumed to have emerged subsequent to the A/B speciation event. Thus, the maximum age of these ORFan proteins is proportional to the relative terminal branch length r. (Relative age of ORFans estimation was calculated from my phylogenetic tree work analysis)

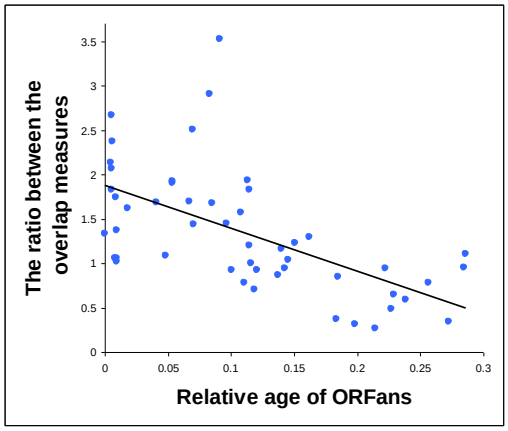
4a. ORFans versus Regular proteins



4b. ORFans versus Random proteins

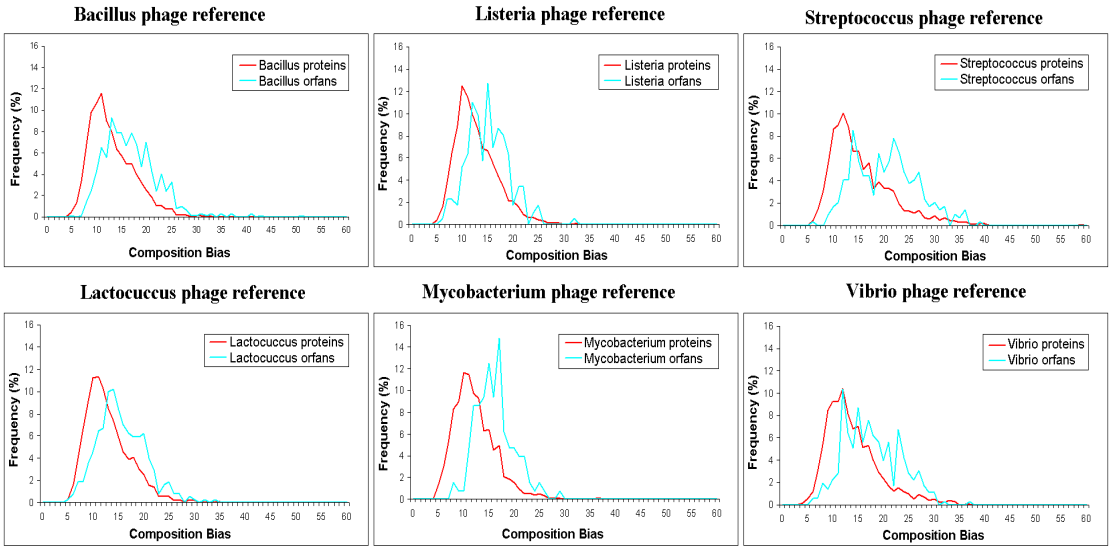


4c. The ratio measure



Supplementary Figure S4: The correlation between the relative age of the ORFans and various measures related to their composition bias. For each organism, the figure shows the correlation between the relative age of its ORFans and the difference between the average composition bias of the ORFans and either real proteins (4a), or random proteins (4b). In both cases there is a significant correlation of -0.66 and 0.59, respectively. (4c) shows similar results by using the correlation between the relative age and the ratio of the overlap of the ORFans and real protein and the overlap of the ORFans and the random proteins (See Section 5

Methods) This ratio reflects the relatedness between the ORFan proteins to either the real proteins (low ratio values) or the random proteins (high ratio values). A correlation coefficient of -0.58 was found. (Relationship of Relative age of ORFans with three different measurements in this result was part of my analysis work.)



Supplementary Figure S5: We selected six bacteria (Bacillus subtilis, Listeria innocua, Lactococcus lactis, Mycobacterium leprae, Streptococcus pneumoniae, and Vibrio cholerae) for which several phages have been sequenced. For each bacterium, we created a set of all the corresponding phage proteins and used it to calculate the composition bias of its bacteriophages. For these six bacteria the averaged composition vector of all the phages associated with each bacterium was calculated. Then, the composition bias of all the proteins (red) and the ORFan proteins (blue) were calculated using the bacteriophage as reference. The results show that the composition of the ORFan proteins in each organism is dissimilar to that of the bacteriophage. (This work has been done by Inbal Yomtovian and Dr. Ron Unger)

Supplementary Table S1. List of genomes studied. Sequences were taken in July 2008 from (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/)

Full Name	Type	Number of Proteins	Number of ORFans	Relative age of ORFans
Aeropyrum pernix	Archaea	1700	85	0.27246
Archaeoglobus fulgidus	Archaea	2420	269	0.25659
Aquifex aeolicus	Bacteria	1529	69	0.22724
Agrobacterium tumefaciens str C58	Bacteria	2765	135	0.08291
Borrelia burgdorferi	Bacteria	851	124	0.19832
Bacillus halodurans	Bacteria	4066	285	0.11475
Brucella melitensis	Bacteria	2059	96	0.09113
Bacillus subtilis	Bacteria	4103	240	0.10794
Clostridium acetobutylicum	Bacteria	3672	438	0.12082
Caulobacter crescentus	Bacteria	3737	340	0.15084
Corynebacterium glutamicum	Bacteria	2993	138	0.13731
Campylobacter jejuni	Bacteria	1634	113	0.13967
Clostridium perfringens	Bacteria	2660	269	0.11886
Chlamydophila pneumoniae AR39	Bacteria	1112	40	0.00013
Chlamydia muridarum	Bacteria	904	55	0.06978
Deinococcus radiodurans	Bacteria	2629	455	0.22898
Escherichia coli O157 H7	Bacteria	5230	133	0.00882
Halobacterium sp NRC 1	Archaea	2075	152	0.28623
Helicobacter pylori 26695	Bacteria	1576	40	0.00932
Listeria innocua	Bacteria	2968	46	0.00817
Lactococcus lactis subsp lactis	Bacteria	2321	144	0.10036
Listeria monocytogenes EGD e	Bacteria	2846	51	0.00917
Methanococcus jannaschii	Archaea	1729	231	0.23841
Mycobacterium leprae	Bacteria	1605	92	0.04065
Mesorhizobium loti	Bacteria	6743	517	0.09650
Mycoplasma pulmonis	Bacteria	782	179	0.21454
Mycobacterium tuberculosis	Bacteria	3989	34	0.00590
Neisseria meningitides	Bacteria	2011	158	0.00507
Nostoc sp PCC 7120	Bacteria	5366	440	0.11484
Pseudomonas aeruginosa	Bacteria	5568	218	0.14244
Pyrobaculum aerophilum	Archaea	2605	424	0.28444
Pasteurella multocida	Bacteria	2015	33	0.05332
Pyrococcus horikoshii	Archaea	1955	73	0.04826
Rickettsia conorii	Bacteria	1374	277	0.01802
Ralstonia solanacearum	Bacteria	3440	158	0.14499
Streptococcus pneumoniae	Bacteria	2105	197	0.06680
Streptococcus pyogenes	Bacteria	1697	112	0.07036
Sulfolobus solfataricus	Archaea	2977	210	0.11047
Sulfolobus tokodaii	Archaea	2825	270	0.11595
Salmonella typhimurium LT2	Bacteria	4425	25	0.00465
Thermoplasma acidophilum	Archaea	1482	50	0.08494
Thermotoga maritime	Bacteria	1858	104	0.22227
Treponema pallidum	Bacteria	1036	223	0.18505
Ureaplasma urealyticum	Bacteria	614	127	0.18326
Vibrio cholerae	Bacteria	2742	137	0.11300
Xylella fastidiosa	Bacteria	2766	626	0.16211
Yersinia pestis	Bacteria	3885	180	0.05344

Bibliography

1. Chothia, C. (1992). "Proteins. One thousand families for the molecular biologist." Nature 357(6379): 543-4.
2. Andreeva, A., Howorth, D., et al. (2008). "Data growth and its impact on the SCOP database: new developments." Nucleic Acids Res 36(Database issue): D419-25.
3. Berman, H.M., Westbrook, J., et al. (2000). "The Protein Data Bank." Nucleic Acids Res. 28(1): 235-42.
4. Berman H.M. (2008). “The Protein Data Bank:a historical perspective”. Acta. Cryst., A64, 88-95.
5. Coulson, A.F. and Moult, J (2002). "A unifold, mesofold, and superfold model of protein fold use." Proteins 46(1): 61-71.
6. Fischer, D. and Eisenberg, D. (1999). "Finding families for genomic ORFans." Bioinformatics 15(9): 759-62.
7. Yin, Y. and Fischer, D. (2006). "On the origin of microbial ORFans: quantifying the strength of the evidence for viral lateral transfer." BMC Evol Biol 6: 63.
8. Yan, Y. and Moult, J. (2005). "Protein family clustering for structural genomics." J Mol Biol 353(3): 744-59.
9. Toll-Riera M, et al. (2009) “Origin of primate orphan genes: a comparative genomics approach”. Mol Biol Evol., 26,603–612.
10. Ekman, D. and Elofsson, A. (2010). “Identifying and Quantifying Orphan Protein Sequences in Fungi”. J Mol Biol., 396 (2), 396-405.

11. Domazet-Loso T. and Tautz D. (2003). “An Evolutionary Analysis of Orphan Genes in *Drosophila”.* Genome Res., 13, 2213 – 2219.
12. Boyer M. et al. (2010)., “Classification and determination of possible origins of ORFans through analysis of nucleocytoplasmic large DNA viruses. Intervirology, 53, 310-320.
13. Haleh Amiri, Wagied Davids, and Siv G. E. Andersson., (2003). “Birth and Death of Orphan Genes in Rickettsia”. Mol Bio Evol, 20(10), 1575-1587
14. Siew,N. and Fischer,D. (2004). “Structural biology sheds light on the puzzle of genomic ORFans.” J Mol Biol. 342, 369-373.
15. Logsdon, J. M., Jr. and W. F. Doolittle (1997). "Origin of antifreeze protein genes: a cool tale in molecular evolution." Proc Natl Acad Sci U S A. 94(8), 3485-7.
16. Wang, W., H. Zheng, et al. (2005). "Origin and evolution of new exons in rodents." Genome Res 15(9), 1258-64.
17. Levine MT. et al. (2006). “Novel gene derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biases expression”. Proc Natl Acad Sci U S A. 103(26), 9935-9.
18. Delaye L, Deluna A, Lazcano A and Becerra A. (2008) “The origin of a novel gene through overprinting in Escherichia coli.” BMC Evol Biol. 28(8), 31.
19. Long,M., Betrán,E., Thornton,K. and Wang,W. (2003) The origin of new genes: glimpses from the young and old. Nat Rev Genet. 4, 865-875.
20. Patthy, L. (2003). "Modular assembly of genes and the evolution of new functions." Genetica 118(2-3), 217-31.

21. Bornberg-Bauer, E., F. Beaussart, et al. (2005). "The evolution of domain arrangements in proteins and interaction networks." Cell Mol Life Sci. 62(4): 435-45.
22. Cortez, D. Q., Forterre, P. and Gribaldo, S. (2009) “A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes.” Genome Biology, 10, R65.
23. Wilson, A. C., Carlson, S.S., and White, T.J. (1977). “Biochemical evolution.” Annu.Rev. Biochem. 46, 573-639.
24. Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. (1978) "A model of evolutionary change in proteins." In "Atlas of Protein Sequence and Structure” 5(3) M.O. Dayhoff (ed.), pp. 345-352.
25. Page R.D.M (2011). “Space, time, form: viewing the Tree of Life.” Trends in Ecology and Evolutionary informatics. 27(2), 113-120.
26. Gogarten, J. P., Doolittle, W. F., Lawrence, J. G. (2002) “Prokaryotic evolution in light of gene transfer.” Mol Biol Evol 19, 2226–38.
27. Battistuzzi, F. et al. (2004) “A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land.” BMC Evolutionary Biology. 4:44.
28. Delsuc, F., Brinkmann, H., and Philippe, H. (2005) “Phylogenomics and the reconstruction of the tree of life.” Nature review. 6: 361-375.
29. Woese, C.R, Magrum, L.J, and Fox, G.E. (1978) “Archaebacteria”. JMol Evol. 74, 5088–99.

30. Hasegawa M, Hashimoto. (1993).”Ribosomal RNA trees misleading?” Nature., 361, 23.

31. Lockhart PJ, Howe CJ, Bryant DA, Beanland TJ, Larkum AW. (1992) ”Substitutional bias confounds inference of cyanelle origins from sequence data”. J Mol Evol., 34**,**153-162.

32. Ludwig W, Klenk H-P.(2000) “Overview: A phylogenetic backbone and taxonomic framework for procaryotic systematic”. In Bergey's Manual of Systematic Bacteriology Volume 1, 2nd edition. Edited by: Boone DR, Castenholz RW, Garrity GM. New York, NY: Springer- Verlag,49-65.

33. Wolf, Y. I., Rogozin, I. B., Grishin, N. V., and Koonin, E. V. (2002) “Genome trees and the tree of life.” Trends Genet. 18,472-479.

34. Ciccarelli, F.D. et al. (2006) “Toward Automatic Reconstruction of highly resolved tree of life.” Science. 311, 1283 -1287.

35. Wu M. and Eisen J.A.(2008).”A simple, fast, and accurate method of phylogenomic inference.” Genome Biology. 9,R151.

36. Wu M. et al. (2009).”A phylogeny-driven genomic encyclopedia of Bacteria and Archea.” Nature. 462,1056-1060.

37. Natalya Y., Pere P, Koonin E.V., Wolf Y.I. (2012).” Phylogenomics of Prokaryotic Ribosomal Proteins”. Plos One. 7(5),e36972.

38. Yan, Y. and Moult, J. (2005) “Computational analyses of microbial genomes: operons, protein families and lateral gene transfer”**.** Online at http://drum.lib.umd.edu/bitstream/1903/2596/1/umi-umd-2490.pdf (as of August 2005).

39. Davies, J. (1996) “Origins and evolution of antibiotic resistance.” Microbiologia.12, 9-16.

40. Lawrence, J. G. and Ochman, H. (1998) “Molecular archaeology of the *Escherichia coli* genome.” Proc Natl Acad Sci U S A. 95, 9413–7.

41. Canchaya, C., Fournous, G., Brussow, H. (2004) “The impact of prophages on bacterial chromosomes.” Mol Microbiol 53, 9–18.

42. Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M., Br¨ussow, H. (2003) Phage as agents of lateral gene transfer. *Curr Opin Microbiol* 6, 417–24.

43. Hendrix, R.W., Smith, M.C., Burns, R.N., Ford, M.E. and Hatfull, G.F. (1999) “Evolutionary relationships among diverse bacteriophages and prophages: all the world’s a phage.” Proc. Natl. Acad. Sci. USA. 96**,** 2192-2197.

44. Yim, G. (April 2009). “Attack of the superbugs: antibiotic resistance”. on The science creative quarterly. http://www.scq.ubc.ca/attack-of-the-superbugs-antibiotic-resistance (accessed June 23, 2009).

45. Osborn, A. M. and Boltner, D. (2002) “When phage, plasmids, and transposons collide: Genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum.” Plasmid 48, 202–12.

46. Gogarten, J. P. and Townsend, J. P. (2005) “Horizontal gene transfer, genome innovation and evolution.” Nat Rev Microbiol 3, 679–87.

47. Frost, L. S., Leplae, R., Summers, A. O., Toussaint, A. (2005) “Mobile genetic elements: the agents of open source evolution.” Nature Rev. Microbiol. 3, 722–732.

48. Nelson, K. E., Clayton, R. A., Gill, S. R.,Gwinn, M. L., Dodson, R. J., Haft, D. H.,Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., Mcdonald, L., Utterback, T. R., Malek, J. A., Linher, K. D., Garrett, M. M., Stewart, A. M., Cotton, M. D., Pratt, M. S., Phillips, C. A., Richardson, D., Heidelberg, J., Sutton, G. G., Fleischmann, R. D., Eisen, J. A., White, O., Salzberg, S. L., Smith, H. O., Venter, J. C., Fraser, C. M. (1999) "Evidence for lateral gene transfer between archaea and bacteria from the genome sequence of *Thermotoga maritima."* Nature 399, 323–9
49. Ochman, H., Lawrence, J. G., Groisman, E. A. (2000) "Lateral gene transfer and the nature of bacterial innovation." *Nature* 405, 299–304.
50. Koonin, E.V., Makarova, K.S., Aravind, L. (2001). "Horizontal gene transfer in prokaryotes: quantification and classification." Annu. Rev. Microbiol. **55**, 709–742.
51. Albà MM, Castresana J (2005). "Inverse relationship between evolutionary rate and age of mammalian genes". Mol Biol Evol 22, 598–606.
52. Albà, M.M., and Castresana, J. (2007). "On homology searches by protein Blast and the characterization of the age of genes". BMC Evolutionary Biology 7, 53.
53. Luz H. and Vingron M. (2005)." Family specific rates of protein evolution." Bioinformatics. 22(10), 1166-1171.
54. Luz H., Staub E. and Vingron M. (2006) "About the interrelation of evolutionary rate and protein age." 17(1), 240-250.

55. Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ (2009). “The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of diffent apparent ages.” Proc Natl Acad Sci U S A 106: 7273– 7280.
56. Vishnoi A, Kryazhimskiy S, Bazykin GA, Hannenhalli S, Plotkin JB (2010) “Young proteins experience more variable selection pressures than old proteins.” Genome Res 20: 1574–1581.
57. Toll-Riera M. et al. (2012).” Origin of primate orphan: A comparative genomic approach. Mol Bio Evol. 26 (3): 603-613.
58. Saeed RT. And Dean C.M. (2006) “Protein protein interactions, evolutionary rate, abundance and age”. BMC Bioinformatics. 7(128), 1 -13.
59. Pal C., Papp B. and Lercher MJ. (2006).” An integrated view of protein evolution.” Nat Rev Genet., 7, 337-348.
60. Plata G., Gottesman M.E. and Vitkup D. (2010). “The rate of the molecular clock and the cost of gratuitous protein synthesis”. Genome Biology, 11, R98.
61. Drummond, D. A., Bloom, J. D., et al. (2005). "Why highly expressed proteins evolve slowly." Proc Natl Acad Sci U S A. 102(40): 14338-43.
62. Drummond CA. and Wike CO. (2008)” Mistranslation-induced protein misfolding as dominant constrain on coding-sequence evolution.” Cell, 134,341-352.
63. Drummond CA. and Wike CO. 2009 “ The evolutionary consequences of erroneas protein synthesis. Nat Rev Genet, 10, 715-724.

64. Yang, Z. 1996. “Among-site rate variation and its impact on phylogenetic analyses”. Trends Ecol. Evol. 11:367–370.
65. Dickerson R.E. (1971).” The structure of cytochrome c and the rates of molecular evolution.” J Mol Evol 1, 26–45.
66. Kim WK, Marcotte EM (2008) Age-dependent evolution of the yeast protein interaction network uggests a limited role of gene duplication and divergence. PLoS Comput Biol 4: e1000232.
67. Blow, N., (2009). Proteins and proteomics: life on the surface, Nature Methods, 6, 389-393.
68. Cai JJ, Woo PCY, Lau SKP, Smith DK, Yuen K (2006) Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in ascomycota. J Mol Evol 63: 1–11.
69. Kunin V., Pereira-Leal J.B. and Ouzounisn C.A. (2004). Functional Evolution of the yeast Protein interaction network. Mol Biol Evol, 21 (7), 1171-1176.
70. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, et al. (2007) Intrinsic disorder and functional proteomics. Biophys J 92: 1439–1456.
71. Dunker A.K., Silman I., Uversky V.N., and Sussman J.L. (2008). “Function and structure of inherently disordered proteins”.Structural Biology, 18, 756 - 764.
72. Fong J.H. et al. (2009) "Intrinsic Disorder in Protein Interactions: Insights From a Comprehensive Structural Analysis. PLoS Computational Biology, 5(3), e1000316.

73. Ward J.J. et al. (2004). “Prediction and functional analysis of native disorder in proteins from the three kingdoms of life.” JMB. 337: 635-645.
74. Meszaros, B. et al. (2009). “Prediction of protein binding regions in disordered proteins. PLoS Comput Biol 5(5), e1000376.
75. Kunihiro B., Shibata R. and Sibuya M. (2004). “Partial correlation and conditional correlation as measures of conditional independence”. Australian and New Zealand journal of Statistics. 46(4), 657-664.
76. Pe'er, I., Felder, C.E., Man,O., Silman,I., Sussman,J.L. and Beckmann,J.S. (2004). “Proteomic signatures: amino acid and oligopeptide compositions differentiate among phyla.” Proteins 54, 20-40.
77. Ermolaeva MD. (2001). "Synonymous codon usage in bacteria". Curr Issues Mol Biol 3(4), 91–7.
78. Suzuki H, Brown CJ, Forney LJ, Top EM (2008). "Comparison of correspondence analysis methods for synonymous codon usage in bacteria". DNA Res. 15 (6): 357–65.
79. Lynn DJ, Singer GA, Hickey DA (2002). "Synonymous codon usage is subject to selection in thermophilic bacteria". Nucleic Acids Res. 30 (19): 4272–7.
80. Paul S, Bag SK, Das S, Harvill ET, Dutta C (2008). "Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes". Genome Biol. 9 (4): R70.
81. Koonin EV, Mushegian AR, Galperin MY, Walker DR. (1997) “Comparison of archaeal and bacterial genomes: computer analysis of protein sequences

predicts novel functions and suggests a chimeric origin for the archaea.” Mol Microbiol., 25:619-637.

82. Karlin, S. (2001) “Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes.” Trends Microbiol 9, 335–43.
83. Kolaczkowski, B; Thornton, JW (2004). "Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous". Nature 431 (701), 980–984.
84. Felsenstein J. (1981).” Evolutionary trees from DNA sequences: a maximum likelihood approach” J.Mol.Evol. 22,160-174.
85. Guindon, S. and Gascuel, O. (2003) “A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood.” Syst.Biol. 52, 696-704.
86. Holder, M.T. and Lewis, P.O. (2003). “Phylogeny estimation: traditional and Bayesian approaches”. Nat.Rev.Genet.,4,275-284
87. Kolaczkowski, B; Thornton, JW (2004). "Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous". Nature 431 (701), 980–984.
88. Saitou, N. and Nei, M. (1987) “The neighbor-joining method: a new method for reconstructing phylogenetic trees.” Mol Biol Evol. 4(4), 406-25.
89. Williams K.P., Sobral B.W., and Dickerman A.W. (2007).” A Robust species Tree for the Alphaproteobacteria.” Journal of Bacteriology. 189(13), 4578-4586.

90. Gao B., and Gupta R.S. (2007). “Phylogenomic analysis of proteins that are distinctive of *Archaea* and its main subgroups and the origin of methanogenesis.” BMC Genomics. 8,86.
91. Satoko N.**,** Yuichi H.,Tomoyuki S. and Moriya O. (2009). **“**Complex coevolutionary history of symbiotic Bacteroidales bacteria of various protists in the gut of termites”. BMC Evolutionary Biology. **9**,158 -170.
92. Wu, G., Fiser, A., Terkuile, B., Sali, A. & Muller, M. (1999). “Convergent evolution of Trichomonas vaginalis lactate dehydrogenase from malate dehydrogenase.”Proc. Natl Acad. Sci. USA, 96, 6285-6290.
93. Dimmer E.C. et al. (2011) “The UniProt-GO Annotation database in 2011.” Nucleic Acids Res. 40, D565- 70.
94. Edgar, R.C. (2004) “MUSCLE:multiple sequence alignment with high accuracy and high throughput.” Nucleic Acids Res. 32(5), 1729-1797.
95. Felsenstein, J. (1989). "PHYLIP - Phylogeny Interface Package (version 3.2)." Cladistics 5: 164 - 166.
96. Jones, D. T., Taylor, W. R., et al. (1992) "The rapid generation of mutation data matrices from protein sequences." Comput Appl Biosci. 8(3): 275-82.
97. Rousseeuw, P.J. and Leroy, A. M. (1987) “Robust regression and outlier detection.” New York, Wiley.
98. Dallal, G.E. and Rousseeuw, P.J. (1992). “LMSMVE: a program for least median of squares regression and robust distances.” Comput Biomed Res 25, 384-91.

99. Parzen, E. (1962) “On estimation of a probability density function and mode”, Ann. Math. Stat. 33, pp. 1065–1076.

100. Massart D.L. and Kaufman L. (1986). “Least median of squares: A robust method for outlier and model error detection in regression and calibration”. 187:171-179.

101. Gentleman R. and Ihaka R. (2000), “Lexical Scope and Statistical Computing: R: A Language for Data Analysis and Graphics”, Journal of Computational and Graphical Statistics, 9, 491–508.

102. Silverman, B. W. (1986). “Density estimation for statistics and data analysis.” London: Chapman and Hall, 48.

103. Felsenstein J (1985).” Confidence limits on phylogenies: An approach using the bootstrap.” Evolution 39,783-791.

104. Touchon M. et al. (2009). “Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths”. PLPS Genetics, 5(1), 1 -25.

105. Vogel C., Bashton M., Kerrison N.D., Chothia C. and Teichmann S.A. (2004). “Structure, function and evolution of multidomain proteins”. Current Opinion in Structural Biology, 14, 208–216.

106. Vitkup D, Melamud E, Moult J, Sander C. (2001). “Completeness in structural genomics”. Nat Struct Biol., 8(6), 559-66.

107. Su C., Peregrine-Alvarez J.M., Butland G., Phanse S., Fong V., Emili A. and Parkinson J.(2008). “Bacteriome.org-an integrated protein interaction database for E.coli.” Nucleic Acids Res., 36, D632-636.

108. Oleg P., Gargac S.M., Cheng Y., Uversky V.N. and Dunker A.K. (2008).”Protein disorder is positively correlated with gene expression in E.coli.” J Proteome Res. 7(6): 2234-2245.

109. Dunker A. and Obradovic Z. (2001). “The protein trinity-linking function and disorder.” Nature Biotechnol. 19, 805-806.

110. Melamud, E. & Moult, J. (2003). Evaluation of disorder predictions in CASP5. Proteins: Struct.Funct. Genet. 53, 561–565.

111. Wei Y et al. (2001). “High-Density Microarray-Mediated gene expression profiling of *Escherichia coli*.” J.Bacteriology 183(2): 545-556.

112. Butland,G., Peregrin-Alvarez,J.M., Li,J., Yang,W., Yang,X., Canadien,V., Starostine,A., Richards,D., Beattie,B. et al. (2005) Interaction network containing conserved and essential protein complexes in Escherichia coli. Nature, 433, 531–537.

113. Soper, H.E., Young, A.W., Cave, B.M., Lee, A., Pearson, K. (1917). "On the distribution of the correlation coefficient in small samples. Appendix II to the papers of "Student" and R. A. Fisher. A co-operative study", Biometrika, 11, 328-413.

114. Kendall, M. (1938). "A New Measure of Rank Correlation". Biometrika, 30 (1–2),81–89.

115. Daubin, V. and Ochman H. (2004) “Bacterial genomes as new gene homes: The genealogy of ORFans in *E. coli*.” *Genome Res.* 14,1036-1042

116. Siew, N, Fischer, D. (2003) ”Twenty thousand ORFan microbial protein families for the biologist?” Structure (Camb) 2003, 11(1), 7-9.

117. Nair,R. and Rost,B. (2003). “Better prediction of sub-cellular localization by combining evolutionary and structural information.” Proteins 53, 917-930.

118. Ofran,Y. and Margalit,H. (2006). “Proteins of the same fold and unrelated sequences have similar amino acid composition.” Proteins 64, 275-279.

119. Koonin E.V. (2005). “Orthologs, paralogs, and evolutionary genomics”. Annual Review of Genetics. 39: 309- 338.

120. Li L, Stoeckert C. and Roos D. (2003). “OrthoMCL: identi cation of ortholog groups for eukaryotic genomes”. Genome Research, 13:2178 - 2189.

121. Dufayard J., Duret L., Penel S., Gouy M., Rechenmann F. and Perriere G. (2005). “Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence database”. Bioinformatics, 21:2596 - 2603.

122. Alexeyenko A., Tamas I., Liu G., Sonnhammer ELL. (2006) “Automatic clustering of orthologs and inparalogs shared by multiple proteomes”. Bioinformatics, 22:9 – 15.

123. Gracy, J. & Argos, P. (1998). Automated protein sequence database classification. I. Integration of compositional similarity search, local similarity search, and multiple sequence alignment. Bioinformatics, 14, 164–173.

124. Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A. & Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucl. Acids Res. 26, 320–322.

125. Ohlson, T., Wallner, B. & Elofsson, A. (2004). Profileprofile methods provide improved fold-recognition: a study of different profile–profile alignment methods. Proteins: Struct. Funct. Genet. 57, 188–197.

126. Edgar, R. C. & Sjolander, K. (2004). COACH: profileprofile alignment of protein families using hidden Markov models. Bioinformatics, 20, 1309–1318.

127. Jozefczuk, S., Klie, S., Catchpole, G., Szymanski, J., Cuadros, I.A., Steinhauser, D., Selbig, J. and Willmitzer, L. (2010). " Metabolomic and transcriptomic stress response of Escherchia coli." 6(364), 1-16.

128. Kang Y., Weber K.D., Qiu Y.,Kiley P.J., and Blattner F. R. (2005) "Genome-wide expression analysis indicates that FNR of Escherichia coli K-12 regulates a large number of genes of unknown function." J Bacteriol 187(3), 1135-1160.

129. Lemey P., Salemi M. and Vandamme AM. (2009). "The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing". 2nd editions, UK: Cambridge University Press.

130. Pruitt K.D., Tatusova T., and Maglott D.R. (2007). "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35: D61–D65.

131. Hirst SJ., Alexander N., McHaourab HS., Meiler J. (2011). "RosettaEPR: an integrated tool for protein structure determination from sparse EPR data". *J Struct Biol* 173:506-14.