A DYNAMIC MODEL FOR THE EVOLUTION OF PROTEIN STRUCTURE

BY

GUY TAL

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Bioinformatics
with a concentration in Crop Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Master's Committee:

Professor Gustavo Caetano-Anollés, Chair
Professor Jay Mittenthal
Associate Professor Kaustubh Bhalerao

ABSTRACT

Protein domains are three-dimensional arrangements of atomic structure that are recurrent in the proteomes of organisms. Since the three-dimensional structure of a protein determines its function, it is the fold, much more than the underlying protein sequence and underlying chemistry, that is evolutionarily conserved. We are interested in probing the history of life with these domain structures and glimpsing qualitative changes over time by studying a dynamic model of protein evolution. Using standard phylogenetic methods and a census of protein domain structure in hundreds of genomes, we have reconstructed phylogenetic trees of protein domains, defined using the Structural Classification of Proteins (SCOP), where the nodes are folds or fold superfamilies (FSFs), the character vector for each node is a list of abundances of said fold or FSF across a range of species that spans all three superkingdoms of life, and the character states are linearly polarized by abundance; higher abundance within and among species equates to older structures and determines tree structure.

Here we explore at what rate fold or FSF variants and new folds or FSFs appear in evolution. We also explore what collective model of proteome evolution explains such rates. Briefly, what are the dynamics of change? A set of birth-death differential equations was selected to capture the change of interest, with one set for folds and another for FSFs. The models assume that at any given moment there are a certain number of different folds or FSFs, with various abundances, and as each fold or FSF diversifies there are slight changes in the folds or FSFs, producing fold or FSF variants. Eventually as the variants continue to diversify and change as well, a new fold or FSF is born. Thus, there are two rate parameters in each model: the growth rate of fold or FSF variants and the rate of appearance of new folds or FSFs. The model governs the rate change of the average total abundance of a fold or FSF with time. It is fit to the tree so only those fold or FSF transitions

actually present in the tree are assumed possible in the equations. It assumes a global perspective: the total abundance of a fold or FSF is that of the fold or FSF across all species, not within one organism. This perspective is used to properly discount terms of horizontal transfer in a birth-death model since such a transfer contributes no new folds or FSFs to the net abundance across all organisms.

Our model determines 1) that there is a tight connection between the history of folds and FSFs, 2) that the corresponding transition probabilities to new variants of a fold experienced a sharp increase just as the transition probabilities to new folds experienced a steep decline and 3) that this simultaneous sharp increase and decline is explainable by and consistent with the combinatorial explosion of structural domains, referring to the period of high combination and rearrangement of domains and distribution of these new combinations in novel lineages, and the rise of organismal diversification. Our simulations suggest a picture of the past in which exploration of protein structure space proceeds much like that of a budding field of knowledge: first, coarse grain discoveries are made, followed by fine-grain elaboration of each once the coarse-grain discoveries have been exhausted.

ACKNOWLEDGMENTS

This work was most certainly collaborative. It could not have been accomplished without the help of Simina Maria Boca, Jay E. Mittenthal, and Gustavo Caetano-Anollés who initially developed the birth-death model for protein structure. Professors Mittenthal and Caetano-Anollés also provided invaluable aid in numerous conversations that helped develop a practical approach to solving the model, evaluating its parameters, and interpreting the results. I am grateful to them both for being a great and complementary set of advisors, especially for Professor Mittenthal's mathematical insight and precision, and desire to keep us on track, and Professor Caetano-Anollés' far-reaching bird's eye view and comprehensive knowledge of the field. I am also indebted to Arshan Nasir and Liudmila Yafrmava for sharing data if ever I asked for it, and Minglei Wang for sharing his time and explaining some of the work he had previously done with *nd* calculations. Last, I am especially indebted to my good friend Kai Zhao who helped me implement my tree de-nesting algorithm in Java and, over the years I've known him, deeply expanded my knowledge, interest, and abilities of computation, and my paranoia of computers.

TABLE OF CONTENTS

CHAPTER 1

BACKGROUND

1.1 The Hierarchy of Protein Structure

The basic biological machinery inside living things requires proteins. The machine's three-dimensional structure determines the machine's function. Since structure does determine function it is also more likely to be conserved [1]. Thus, if we understand the evolutionary history of protein structure, we can reconstruct some of the evolutionary history of life itself. This is the inspiration for this research.

Proteins are linear polymers of amino acids. This sequence of amino acids corresponds to a three-dimensional structure that determines the functions a protein can perform. As [2] details, Linderstrøm-Lang and Schellman proposed in the 1950's that protein structure had a four-tiered hierarchy, which they called primary, secondary, tertiary, and quaternary structure. These corresponded, respectively, (i) to the amino acid sequence linked by peptide bonds, (ii) the helices and sheet elements of a fold that arise as a result of the hydrogen bonding patterns, (iii) the molecular fold itself, and (iv) the aggregate of such chains into a larger biological construct from which function arises. Recognition of repeating motifs in protein structure led to the advent of (v) supersecondary structure. The further observation that protein folds have modular components that act alone or with other modules in multi-"domain" proteins forms the basis for (vi) protein domains.

It is with domains that this research begins. There are dozens of domain classification systems, but [2] presents a solid argument for the use of SCOP (Structural Classification of Proteins) which we adhere to. "SCOP domains that are closely related at the sequence level (generally expressing $> 30\%$ pairwise amino acid residue identities) are pooled into fold families (FFs), FFs sharing functional and structural features suggestive of a common evo-

lutionary origin are unified further into fold superfamilies (FSFs), and FSFs that share similarly arranged and topologically connected secondary structures are grouped further into protein folds. Folds are then grouped into protein classes according to organization of secondary structure in the fold, defining the major α/β, $\alpha+\beta$, all-α, all-β, small and multidomain groups" [2]. Note that each classification, beginning with fold families on up to classes, is a subset of the next. This research specifically focuses on folds and FSFs because all relevant pieces, but most significantly a "molecular clock," described below in section 2.5, has been discovered for these two classes.

As explained below, a number of studies, [3] and [4], have constructed phylogenetic trees that describe the evolution of folds and FSFs using parsimony methods and genomic abundance of folds or FSFs respectively as phylogenetic characters, and that use the assumption that the character states are linearly polarized by abundance. The rooting was performed using the Lundberg method which does not require the need of outgroups. The topology of the trees provides a static characterization of the evolution of protein *structure*, a word which we use throughout this paper to mean either folds or FSFs. What is missing is a dynamic model detailing the essential parameters relevant to describing the evolution of such a tree. Such a phenomenological description of the dynamics of the process would specify parameters that later could be interpreted in terms of mechanistic molecular processes. We have provided such a description, building on previous models that attempted to explain the evolution of protein structure.

The models of [5] and [6] aim to explain the existence of the power-law distribution of fold occurrence in genomes in which a few folds occur in many copies per genome and many folds are rare. [5], however, take a "local" perspective, i.e. the focus of their model is the distribution within a *single* genome over time. This does not capture a given fold's history as it develops in many organisms across the entire planet. Moreover their model also assumes that the "rate of fold flow" or (fold acquisition − fold deletion) is the same value for all folds within a given organism and that the initial rates of fold duplications within an organism are also identical. Both assumptions seem implausible as there are different selection pressures on different folds, since different folds relate to different functions. We believe that a model taking a global perspective, with the fold across all organisms as its subject, and allowing the two rates discussed to change as they may, is more realistic.

Similarly, in order for the model of [6] to be compatible with the power law distributions they hope to explain they need to assume that domain families are commonly in a state of equilibrium with respect to the total number of domain families over time and the total number of domain families of a given size in a given genome over time. This latter approach appears implausible as there does not appear to be anything like environmental equilibrium at many time scales throughout history, resulting in constantly changing selection pressures on organisms and therefore on the genetics and proteins they contain and encode. Again, we believe that a model relinquishing the assumption of such an equilibrium can shed light on evolutionary history.

Another approach to explaining the power law distributions of protein structures, in [7], takes an information-theoretic approach. They conclude that the power laws of sequences have different origins than those of folds; "protein sequences exhibit a power law distribution to achieve efficient coding of necessary folds" while that of folds "is based on the thermodynamic stability of folds." But here, too, the focus is on a local model, rather than a global model, so insights and benefits of a global perspective are missed.

Yet another approach, found in a study by [8], attempts to reconstruct protein evolutionary history via simulation. [8] demonstrates that, given stability as the selection mechanism in each iteration, random protein sequences often converge on proteins with specific structures, so called "wonderfolds." The authors consider these emergent thermostable wonderfolds as potential pre-biotic precursors to the biotic protein structure hierarchy, ones that likely arose in a hot environment and needed their thermostability. This study, however, focuses only on the pre-biotic world. In contrast, our study focuses on the rise of the biotic world.

Still another approach in [9] takes the organism, rather than the protein structure, as its point of focus, with organisms evolving in a simulation under the assumption that an organism's fitness corresponds to its proteins' abilities to be in their native conformations. This model, too, successfully recovers known power laws and structural hierarchy, but again does not take our global structure model perspective, and the implications that arise from it. To the best of our knowledge no other group has taken the global model perspective.

Thus, we use the data from the phylogenetic tree reconstructions described to build global, dynamical models for the evolution of folds and FSFs and

evaluate their parameters. In contrast to the preceding models, our models (one for folds and another for FSFs) predict the abundance of individual folds or FSFs throughout the living world and generates a non-steady-state time course for their evolution. This paper describes our models, their evaluation using a large sample of real data from all three superkingdoms of life, and the implications for the history of protein structure.

1.2 A Tree of Protein Structure

[10] and [11] built phylogenetic trees using protein structures as taxa and the number of protein structures appearing in various genomes as characters. Note that there is one tree for FSFs and another for folds. The basic assumption is that the structures which are present in greater numbers at present are older than those present in lower numbers, as the more abundant structures had more time to grow in number. The authors call this abundance-based approach genomic demography. The tree was built using the parsimony reconstruction. It was later expanded to include a greater number of structures and organisms. This study uses and builds upon data from those previous studies.

A given structure has potentially many different proteins that are classified within this structure, and we call these the structure's *variants*. In the tree of protein structures each node corresponds to a structure with a vector of character states, the abundance of variants of each type of structure in a set of genomes. The structure originates at its node of origin, which is an internal node of the tree.

1.3 The Parameters for the Model

We built dynamic models of evolution that correspond to our trees of protein structures. We assume that structures evolved in a stochastic branching process from a primordial structure. We consider the ensemble of possible realizations of the branching process. We model the time course of changes in the ensemble-averaged abundance of each structure. The parameters of the model are transition probabilities per unit time, or rate constants, for

transition to a new variant of a structure and for transition from one kind of structure to another. An approximation technique is presented which allows the calculations of these rates and their changes over time.

CHAPTER 2

THE MODEL

2.1 The Global Model of Fold Evolution

In considering a possible model, one must take into account the ways in which a structure can be lost, copied, or transformed into another structure. A new structure can evolve through the mutation of an existing structure (structure transitions), horizontal transfer from another genome, or *de novo* evolution from a non-coding region. We describe structure transitions in the structure transition matrix: It gives the probability per unit time, per domain, that a structure will be transformed into any other structure. We assume that *de novo* evolution is very improbable, as the probability of an open reading frame (the portion of DNA which is transcribed and can eventually be translated into a protein) evolving from a non-coding region is very low. We ignore this.

However, it is worth noting that it has been demonstrated that, at least in *S. cerevisiae*, *de novo* evolution may not be unlikely through translation of transitory proto-genes in non-genic sequences [12]. Momentarily, it is difficult to account for this term in our model as there is not enough data on how widespread this proto-gene mechanism is across the three superkingdoms, how much similarity can be detected among the proto-genes between species, and whether these proto-genes fall into classifiable three dimensional structures. However, if it turns out that the mechanism is important and data can be collected on the three dimensional structures' abundances among and between species, it may then be possible to repeat our study by creating a tree of structures and proto-structures (corresponding to proto-genes), and proceeding as below.

In any case, a copy of a structure can be lost by domain deletion. There is also a selection bias in the occurrence of structure: If a structure appears

and is not soon useful, it will be lost. We opted to use a global model (which deals with the distribution of structures in all organisms), as opposed to a local model (which deals with the distribution of structures in individual organisms). The global model allows us to avoid using separate terms for horizontal transfer. Moreover, it focuses on the structures which have been successful, thus including selection bias in the structure transition probabilities. The global model is a deterministic model for the ensemble-average global abundance of a structure, not a stochastic model.

2.2 The Governing Equations of the Global Model

We will consider a simple model, which will allow an evaluation of all the parameters from the data. We call this model the *irreversible tree-hugging model (ITH)*. In *ITH* we assume that we only have "forward" transitions. That is, the only transition probabilities that are nonzero are those from an ancestral structure to a neighboring descendant; thus transitions are irreversible. We set the transition probabilities for all "reverse" transitions equal to 0. The model also assumes that the only possible transitions are the ones seen in the tree, thus it is tree-hugging. The governing equations of a complete global model would be:

$$\frac{dN_j}{dt} = (\lambda_j - \sum_{i \neq j} a_{ij})N_j + \sum_{i \neq j} a_{ji}N_i \tag{2.1}$$

where:

$N_j(t)$ the global abundance of structures of type j at time t. $t = 0$ at the origin of the primordial structure ($j = 1$). We assume $N_1(0) = 1$ and $N_j(0) = 0$ for $j > 1$. If structure j is ancestral to structure i, $j < i$.

λ_j is the net rate (birth minus death) per copy of structure j of generating new variants of structure j.

a_{ij} = the rate of transition from structure j to structure i (*forward* transition). Note that $a_{ij} \neq 0$ only when structure i is the neighboring descendant of structure j. In particular, this means that each of the two sums in equation

2.1 only has one non-zero term, yielding:

$$\frac{dN_1}{dt} = \lambda_1 N_1 - a_{2,1} N_1 \tag{2.2}$$

$$\frac{dN_j}{dt} = (\lambda_j - a_{j+1,j}) N_j + a_{j,j-1} N_{j-1} \tag{2.3}$$

2.3 Structure Abundance and Funnels

Each internal node in the tree of protein structure has two child nodes, one of them representing the continuation of the old structure, the other the creation of a new structure. Starting from a particular parent node which represents some structure j, we can follow the path through the branches on the tree, which correspond to structure j up to the present (see Figure 2.1). The number of copies of a particular structure is increasing in an approximately exponential manner, and thus we say it corresponds to a funnel (see Figure 2.2). The number of copies of a particular structure at a certain time is represented by the cross-section of the funnel. Each new branch coming off this path corresponds to the initiation of a new structure j' (see Figure 2.1), and thus of a new funnel (see Figure 2.2). Each new funnel initially starts out with one copy of a new structure. We thus note that all the internal nodes represent present-day structures, even though the examples of the structures may have changed over time due to mutations. It is thus possible to identify each one uniquely using its global abundance: The structure with the greater abundance will be the older branch.

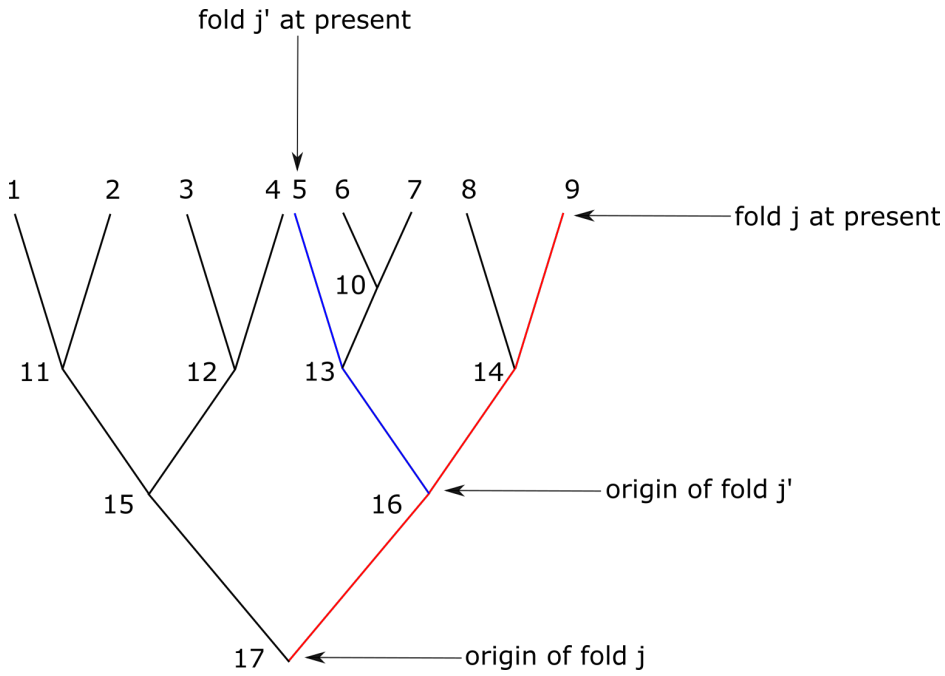

Figure 2.1: Identification of internal nodes

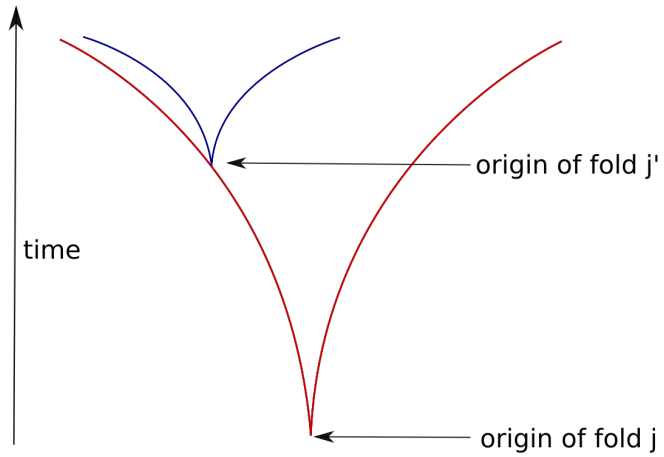

Figure 2.2: The funnels corresponding to folds j and j'. A horizontal cross-section of a funnel measures that funnel's total abundance with respect to time, $N(t)$.

2.4 Estimating the Global Abundances of Structures

Our strategy is to estimate the global abundance of structures, and then to use these estimates to estimate the λ's and a's from N's.

The phylogenetic tree was derived from data for the local abundance of the J folds in K genomes. The relation between the global abundance of structure j, N_j, and the vector of local popularities $[s_{j1}, s_{j2}, ...s_{jK}]$ is

$$N_j = \sum_k m_k s_{jk} \tag{2.4}$$

Here m_k is the effective population size of species k. In words, to calculate the total abundance of a structure, one multiplies the number of times that structure is present in a given species by the number of *breeding* members of that species, and repeats this operation for each species. The *breeding population* size is the intuition behind the formal effective population size which is defined as the size of the ideal population, which acts the same as the actual population. The ideal population assumes no selection, random mating, and a random chance for each offspring to have a particular parent [13]. For ancestral nodes species k will not in fact be the present species k, but rather the lineage leading to the present species k. Thus we are not saying that species k existed in the remote past.

We have data for the present abundance vectors. We require a method of estimating present effective population sizes, denoted by m_k^* for species k. Figure 1B in [14] offers the relationship $\log m_k^* \mu = -1.3 - 0.55 \log G_k$ where G_k is the genome size of species k in millions of bases, Mb. μ_k represents, for species k, mutations/base/cell division. Cell division corresponds to genome duplication for unicellular organisms such as prokaryotes and lower eukaryotes. For multicellular organisms such as higher eukaryotes $\mu_k = \mu_{bs}/c$ where μ_{bs} = mutation rate/base/generation from Figure 3 in [15], and c = number of germline cell divisions [14]. However, [14] also references an average μ_k, which we call μ, and offers its value as $\mu = 2.3 * 10^{-10}$. Using this value, we can solve for m_k^*:

$$m_k^* = 2.2 * 10^8 * G_k^{-0.55} \tag{2.5}$$

2.5 The Issue of a Time Scale

We do not know how to directly assign times to the origins of most structures. However, [16] graphs the geological times of known structures, determined from independent archaeological evidence, against their normalized distance in nodes (nd) from the hypothetical ancestral structure at the base of a phylogenetic tree of structures. This relationship is linear, is presented in equations 2.6 and 2.7, where t is in gigayears, and is assumed to hold for all structures in the study.

Determining the number of bifurcations since the root to any structure can be done by counting. For example, in Figure 2.1, there have been three bifurcations that lead to the birth of leaf 1, but only two bifurcations that lead to leaf 3. The maximum number of bifurcations that lead to a structure on the tree is 3, and we call this the *normalization.* When one takes the number of bifurcations leading to a particular structure and divides by the normalization one arrives at the *normalized node distance* (nd) of a node. The relation between t and nd for folds, as given in [16], was:

$$t = -3.8023nd + 3.8137 \tag{2.6}$$

While for FSFs it was:

$$t = -3.8314nd + 3.6284 \tag{2.7}$$

We assumed these relations hold for all nodes of the respective trees. Since we can calculate λ's, a's, and nd's for folds and FSFs, we can use the above equations to relate our results to real time and make statements about the history of protein structure.

2.6 Estimating λ's

We assume that transitions to new structures are much rarer than transitions to variants, so $a_{j'j} \ll \lambda_j$. Applying this to equations 2.2 and 2.3 it follows that, for a given node h, $N_h = v_{h-1} * e^{\lambda_h(t_n - t_{n-1})}$ where $v_{h-1} = 1$ if node h represents a new structure originating from the structure at node $h-1$, and $v_{h-1} = N_{h-1} - 1 \approx N_{h-1}$ if the structure at node h is the same as the

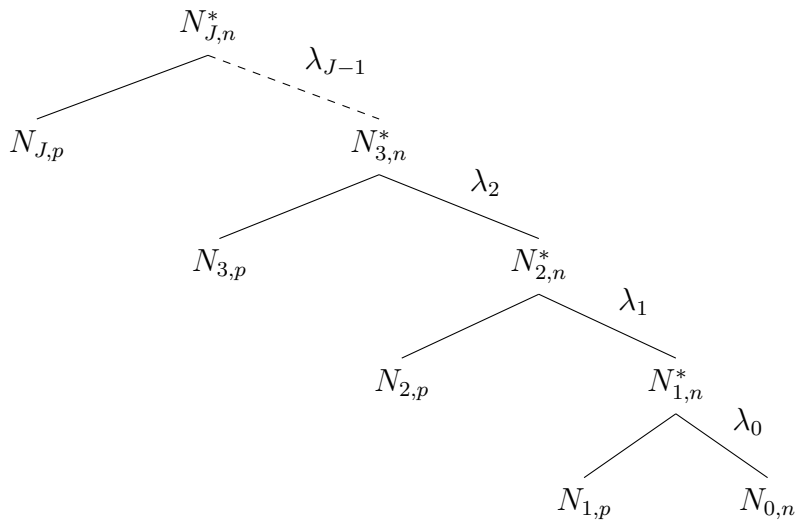

Figure 2.3: A p-comb

structure at node $h-1$. If the node represents a new structure, we call it a *novel structure*, and if the node is the same we say it is a *persistent structure*. This approximation is reasonable because $N_{h-1} \gg 1$ for a persistent structure if $a_{j'j} \ll \lambda_j$, as our results show.

In any case, a novel structure requires knowledge of the present total abundance and time to determine λ while a persistent structure requires, in addition, the total abundance one tree step back. As such, we are faced with several difficulties: 1) determining which structures are novel and which are persistent; 2) determining ancient abundances for persistent structures. Since both of these tasks are difficult to do correctly in principle without making unsavory assumptions, we instead resort to an approximation.

Note that every n-furcation on a tree contains exactly one persistent structure. The rest are novel. Thus, there are numerous labelings of novelty and persistence on a tree to consider in search of the historically accurate one. We consider two interesting extremes which we call a *p-comb* and an *n-comb*. A p-comb, as illustrated in Figure 2.3, is a binary tree in which, at each bifurcation, the persistent structure becomes a leaf (as witnessed by the second subscript p labels in the Figure and equations). An n-comb, as illustrated in Figure 2.4, is a binary tree in which, at each bifurcation, the novel structure becomes a leaf (as witnessed by the second subscript n labels in the Figure and equations).

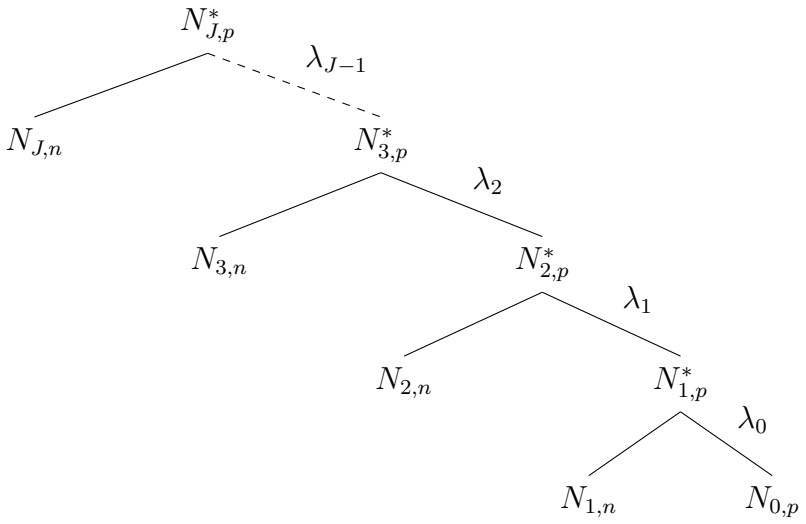

Figure 2.4: An n-comb

For the p-comb the following equations follow from the above discussion.

$$\begin{aligned} N_{0,n} &= e^{\lambda_0 t} \\ N_{1,p} &= N^*_{1,n} e^{\lambda_1 t} \\ N^*_{1,n} &= e^{\lambda_1 t} \\ N_{1,p} &= e^{2\lambda_1 t} \end{aligned}$$

where the last equation follows from the previous two and the first equation gives $\lambda_{0,n}$ directly. Similarly,

$$\begin{aligned} N_{2,p} &= N^*_{2,n} e^{\lambda_2 (2t)} \\ N^*_{2,n} &= e^{\lambda_2 t} \\ N_{2,p} &= e^{3\lambda_2 t} \end{aligned}$$

Note here that t has been replaced by $2t$ in the first equation since this structure has been around for that length of time since its origin. It follows

by a generalization of the above that:

$$N_{j,p} = e^{(j+1)\lambda_j t} \tag{2.8}$$

$$\lambda_{j,p} = \frac{\ln N_j}{(j+1)t} \tag{2.9}$$

$$\lambda_{0,n} = \frac{\ln N_0}{t} \tag{2.10}$$

For the n-comb, however, the calculation is different:

$$\begin{aligned} N_{0,p} &= N^*_{1,p} e^{\lambda_0 t} & \qquad N_{1,n} &= e^{\lambda_1 t} \\ N^*_{1,p} &= N^*_{2,p} e^{\lambda_0 t} & \qquad N_{2,n} &= e^{\lambda_2 (2t)} \\ N^*_{2,p} &= N^*_{3,p} e^{\lambda_0 t} & \qquad N_{3,n} &= e^{\lambda_2 (3t)} \end{aligned}$$

Clearly, the left-hand side of the above is a chain of equations that can be easily solved while the right-hand side can be generalized to:

$$N^*_{0,p} = N^*_{J,p} e^{\lambda_0 t} \qquad\qquad N_{j,n} = e^{\lambda_j (jt)}$$

This can be solved to yield:

$$\lambda_{j,n} = \frac{\ln N_j}{jt} \tag{2.11}$$

$$\lambda_{0,n} = \frac{\ln \frac{N_0}{N_j}}{jt} \tag{2.12}$$

Notice that combining our analysis for p-comb and n-comb trees, we get the ratio:

$$\frac{\lambda_{j,n}}{\lambda_{j,p}} = \frac{j+1}{j}, \ j = 1, ..., J \tag{2.13}$$

This latter suggests that whether we are dealing with an n-comb or a p-comb the calculations for λ's will not differ by much. We say that these two extremes *bracket* the space of real solutions, though we do not know whether they bound the space.

If we imagine a random walk in sequence space, one interpretation for a p-comb is that new structures arise at the boundary of a region for a previously

new structure. An n-comb, on the other hand, implies that the first structure is the stem line and its boundary contacts regions for all other structures.

Clearly, the real tree is neither a p-comb nor an n-comb nor, in fact, binary (though it departs from binary-ness in only a handful of nodes). We will discuss how comb-like our tree is and delve into easing these approximations in 3.4.

2.7 Estimating a's

To derive the a's we use an approximation which we call the "nuclear decay" approximation for reasons that will become clear. Consider structure j during the time interval between its origin and the time which a new structure, j', originates from j. For convenience of notation we take $t = 0$ as the time at which j originated – that is, $N_j(0) = 1$. Suppose transitions to new structure are rare, so $a_{j'j} \ll \lambda_j$. Then approximately, $N_j(t) = e^{\lambda_j t}$. Now, imagine plotting $t(N_j)$ as a logarithmic curve on a plot of N_j (abscissa) vs time (ordinate). To each value of N_j corresponds a value of $\tau(N_j)$, the mean time interval from t to the first transition of a copy of structure j to j'. τ decreases as N_j increases because the more copies of structure j exist the more likely it is that one of them will make the transition to j'. Below we show that $\tau = \frac{1}{N_j^2 a_{j'j}}$. Imagine plotting this function $\tau(N_j)$ above the function $t(N_j)$ so the upper curve shows $t(N_j) + \tau(N_j)$. This curve has a minimum value which is the mean time at which the first transition to j' will occur. These three curves are illustrated in Figure 2.5, with the minimum value visible on the green curve representing $t(N_j) + \tau(N_j)$.

To find an analytic expression for this minimum value, note that the transition $j \to j'$ is analogous to the radioactive decay of an atom, which is described by a Poisson process. Let the probability that an atom decays in the time interval dt be adt. The probability that one of N atoms will first decay between t and $t + dt$ is

(prob that none decay before t)(prob that one decays in dt) = $([e^{-at}]^N)(adt)$

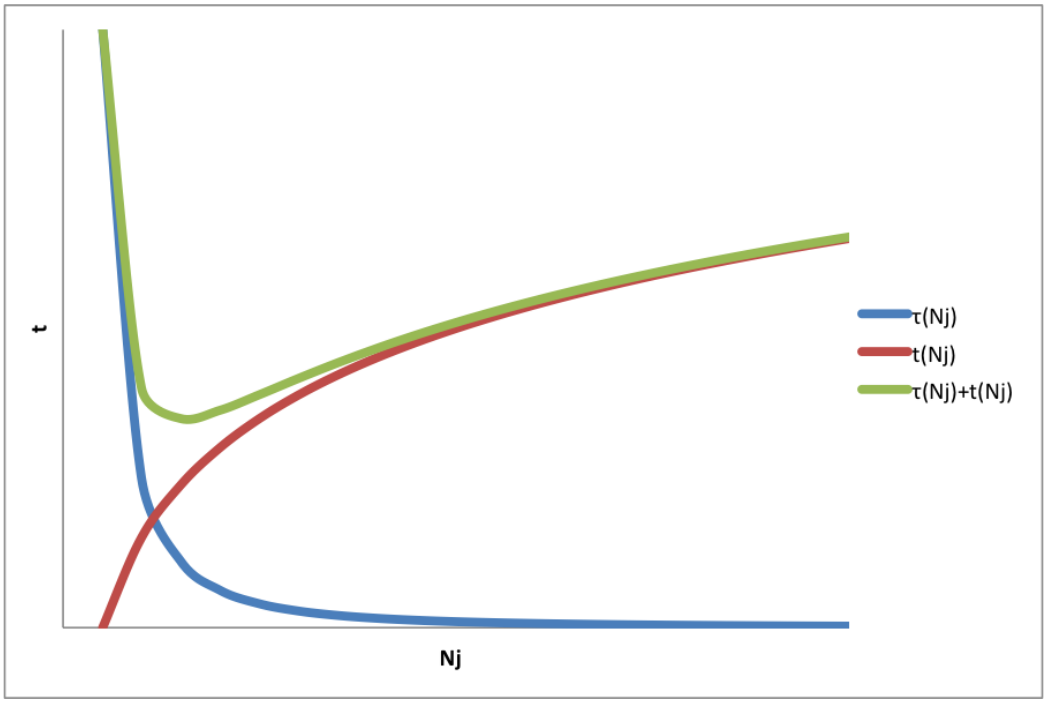

Figure 2.5: Graph of $t(N_j)$, $\tau(N_j)$, and $t(N_j) + \tau(N_j)$

Then the mean first decay time for N atoms is

$$\tau = \int_0^\infty t' e^{-Nat'} a dt'$$

$$\text{so } \tau = \frac{1}{N^2 a}.$$

Now consider the transition $j \to j'$. From the preceding,

$$t = \frac{1}{\lambda_j} \ln N_j \text{ and } \tau = \frac{1}{N_j^2 a_{j'j}}$$

$$\text{so } \frac{d(t+\tau)}{dN_j} = \frac{1}{\lambda_j N_j} - \frac{2}{a_{j'j} N_j^3}$$

$$\text{which is zero at } \frac{a_{j'j}}{\lambda_j} = \frac{2}{N_{jmin}^2} \tag{2.14}$$

Thus the ratio $\frac{a_{j'j}}{\lambda_j}$ can be determined from the value of N_j when the transition occurs. Because $N_{jmin}^2 \gg 1$, the ratio is $\ll 1$, as we assumed. Since the value of λ_j has been determined we can get $a_{j'j}$ from the ratio.

Notice that the above model was explicated for the trees of protein folds

and FSFs but can be adapted without modification for any level of the protein architectural hierarchy for which all the relevant data discussed is available, namely abundance data and a known relationship between nd and time.

CHAPTER 3

MATERIALS AND METHODS

3.1 Phylogenetic Analysis

Two analyses were performed at different levels of the protein structure hierarchy, the first on *folds* and the second on *fold superfamilies* (FSFs.) The data used to construct the trees consisted of fold and FSF abundances. There were 1030 folds in 749 species across all three super kingdoms. There were 1733 FSFs in 981 species across all three super kingdoms. SCOP was used for both fold and FSF classification. These numbers were the latest available at the time of retrieval. An argument for the virtues of SCOP over other classification schemes such as CATH is given in [4]. While we have placed the graphs for folds and FSFs alongside each other in Figures 4.7-4.9, discussed later, to demonstrate the similar evolution of folds and FSFs, the graphs are not directly comparable because of the different numbers of organisms present in each study.

PAUP* was used to construct phylogenetic trees for both FSFs and folds following the previously described methods [4], with genomic abundance within species serving as characters. A single g value for each structure in each species resulted in two matrices, 1733x981 and 1030x749 for FSFs and folds, respectively. Since large genomes are more likely to have larger protein structure abundances, we normalized the abundances (g) to a linearly ordered 0-23 scale using the formula (for FSFs):

$$g_{ab\_norm} = \text{round}[\frac{23\ln(g_{ab}+1)}{(g_{max}+1)}] \tag{3.1}$$

And for folds, on a linearly ordered 0-20 scale:

$$g_{ab\_norm} = \text{round}[\frac{20\ln(g_{ab}+1)}{(g_{max}+1)}] \tag{3.2}$$

Here, g_{ab} is the g value of the FSF (or fold) a in species b. g_{max} is the maximum g value in the matrices above. These normalized matrices were then handed to PAUP* to compute the trees described above in a standard NEXUS file.

3.2 Calculating *nd*

Trees for folds and FSFs were returned in Newick format. In Newick format numbers represent leaves and parentheses around two or more nodes represents an immediate common ancestor among those. For example, $((1,2),(3,4),5)$ corresponds to the tree in Figure 3.1. The leaves of this tree are $1, 2, 3, 4$, and 5 and the internal nodes of the tree are $(1,2),(3,4)$, and $((1,2),(3,4),5)$ representing nodes $6, 7$, and 8 respectively. In order to calculate the *nd* value of any node in a tree given in Newick format one must count the number of left parentheses, ' (', to the left of the node and subtract from this the number of right parentheses, ') ', to the left of the node. For example, for the leaf 3 there are 3 left parentheses to its left and 1 right parentheses to its left, so its unnormalized *nd* value is 2, which is confirmed by inspection. Similarly, the internal node $(3,4)$ has 2 left parentheses and one right parenthesis to its left so it's unnormalized *nd* value is 1, as can again be verified by inspection.

Thus, a Java program was written to de-nest the tree, which is to say get a listing of all its internal nodes and leaves, find the location of each node within the tree's Newick format, and count the number of left and right parentheses to the left of the node to yield the node's *nd* value.

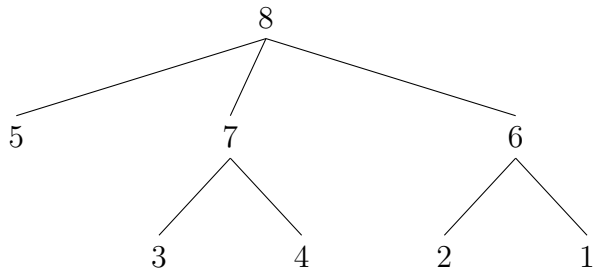

Figure 3.1: A tree corresponding to Newick format ((1,2),(3,4),5)

3.3 Collecting Genomes

Genome sizes for all organisms were mainly downloaded from NCBI, genomes online.org, and project webpages. As a check of sanity, the sum of genome abundances for a given species was graphed against that species' genome size. We expected, and found, a generally increasing trend; a species with a larger genome is more likely to have a larger sum. No extreme or unexpected outliers were found, though the data did have an expected scatter. This data is presented in section 4.1.

3.4 Finding the Nodes in a Comb

The comb analysis, calculating λ's and a's assuming p-combs and n-combs described in chapter 2, was first performed only on the *comb-like* leaves, that is, those leaves that sprang directly from the stem line of the tree. There were 187 such nodes for FSFs and 147 for folds, discovered by manual inspection, accounting for 11% and 14% of the nodes of the respective trees. The analysis was then repeated for all leaves of the trees, assuming each sprang directly from the stem line and the results of the previous calculation were compared. There were no major differences in the results.

Justification for the lack of bias in the comb approach was sought by binning the count of nd for each of the comb nodes; each comb node's nd value was binned in one of a sequence of bins of size 10 spanning the entire range of possible nd values for the structure tree. A uniform distribution was discovered, with one spike at the combinatorial explosion, as one might expect given a look at the spiky nature of the raw data at this point. These results are presented and discussed in 4.2.

3.5 Calculating the Total Abundances, λ's, and a's

As denoted in equation 2.4 the total abundance of a protein structure is given by the dot product of the effective population size vector and the local genomic abundances vector. The effective population size of a particular species is itself dependent on the genome size of that species as noted in equation 2.5, so the effective population size vector is dependent on the corresponding

genome size vector. Note, however, that the exponent on G in equation 2.5 is a derivation from population genetics. We were curious as to how robust our qualitative results were to variations in this exponent and this was explored by repeating the analysis through a range of different exponents from -2 to 2. This procedure was performed in an Excel spreadsheet using array formulas. The λ's and a's were also calculated in this spreadsheet and all results were graphed against time using equations 2.6 and 2.7. Qualitatively, the results presented in this paper appeared to hold well for values of G's exponent in that range, and the overall pattern evinced disappeared in a continuous fashion as we got further away from the exponent's actual value, leading us to believe that our results are robust in the face of a potentially oversimplified population genetics model.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Sum of Genome Abundances vs Genome Size

As discussed in section 3.3, as a check, the sum of fold abundances for a given species was graphed against that species' genome size. Each data point was created by summing the abundances of all folds present in a given species and graphing it vs. that species' genome size. A generally increasing trend was expected and found, as shown in Figures 4.1-4.3, with a species with a larger genome more likely to have a larger sum. Below the data for folds is presented separately by superkingdom because, while the data is increasing for all superkingdoms, for Archaea and Bacteria the graphs are linear while for Eukaryotes a power law is a much better fit. The linear graphs show an abundance of about 1.25 copies per gene.

It's noteworthy that there are known power laws for local structure abundance [5] within a genome, but the sum of these local abundances across species, for a given fold, shows a linear relationship in Bacteria and Archaea, as shown in Figures 4.1 and 4.2. However, the power law is preserved for Eukaryotes. The Metazoan kingdom is the largest contributor to the scatter in this power law. When Metazoa are removed the R^2 noticeably increases as can be seen from Figures 4.4 and 4.5.

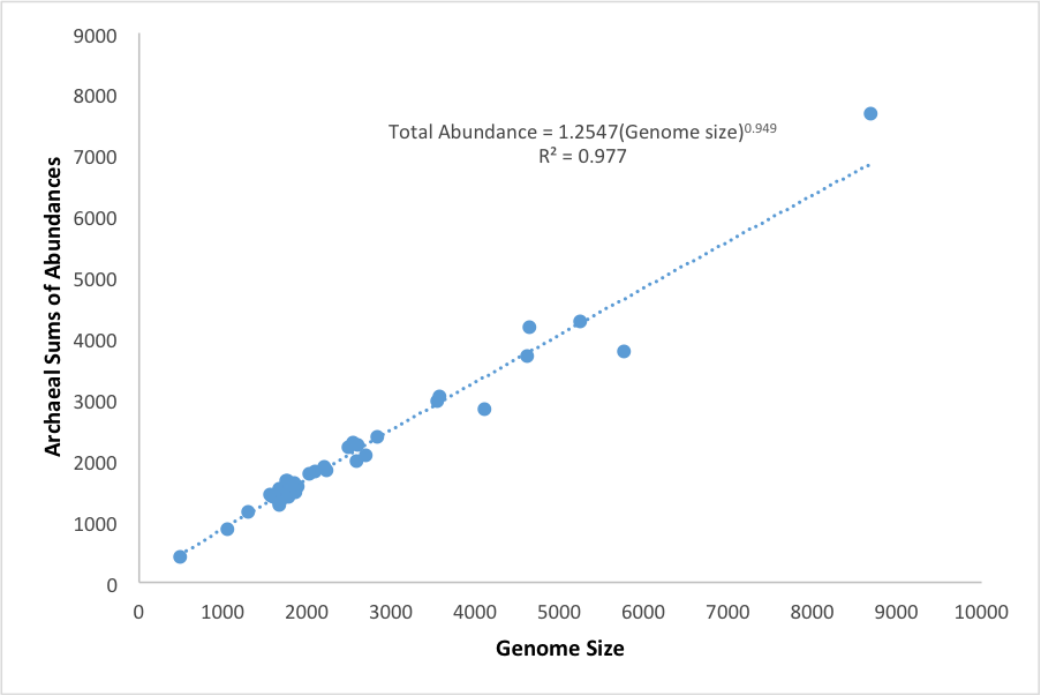

Figure 4.1: Archaeal Sums of Fold Abundances vs Genome Size

Figure 4.2: Bacterial Sums of Fold Abundances vs Genome Size

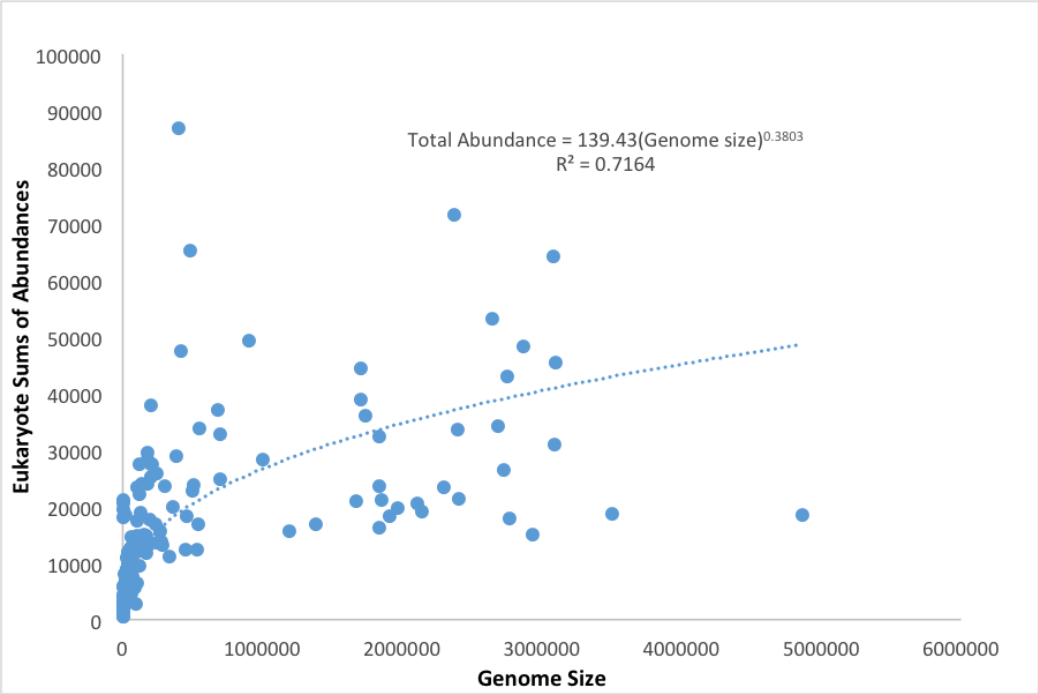

Figure 4.3: Eukaryotic Sums of Fold Abundances vs Genome Size

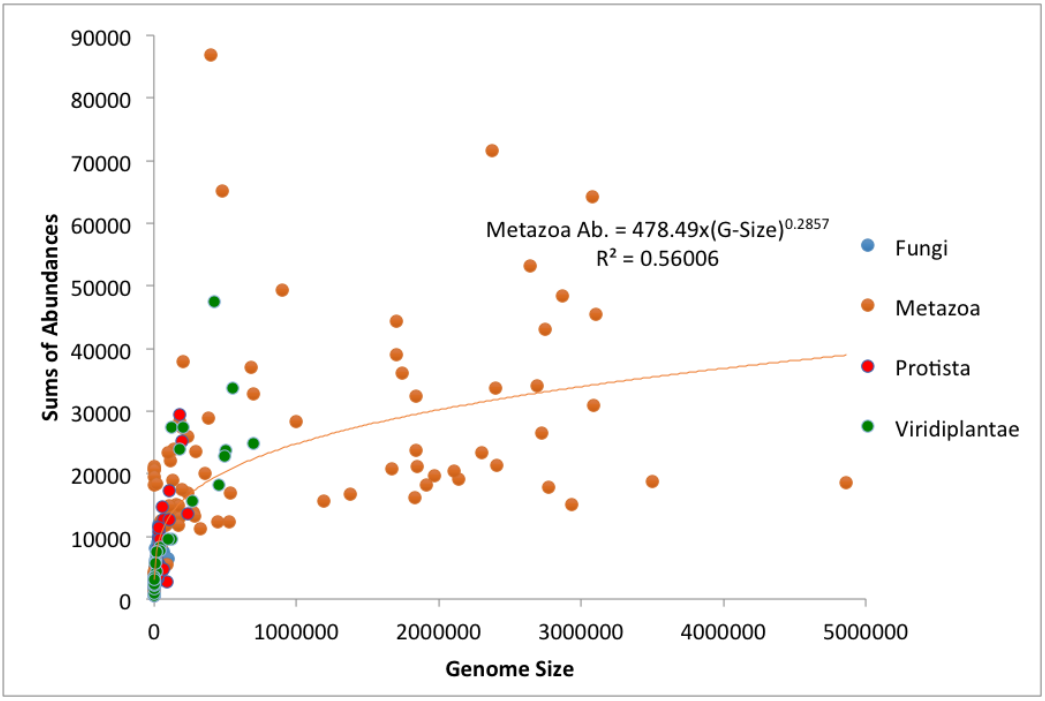

Figure 4.4: Sums of Fold Abundances vs Genome Size for Fungi, Metazoa, Plants, and Protista

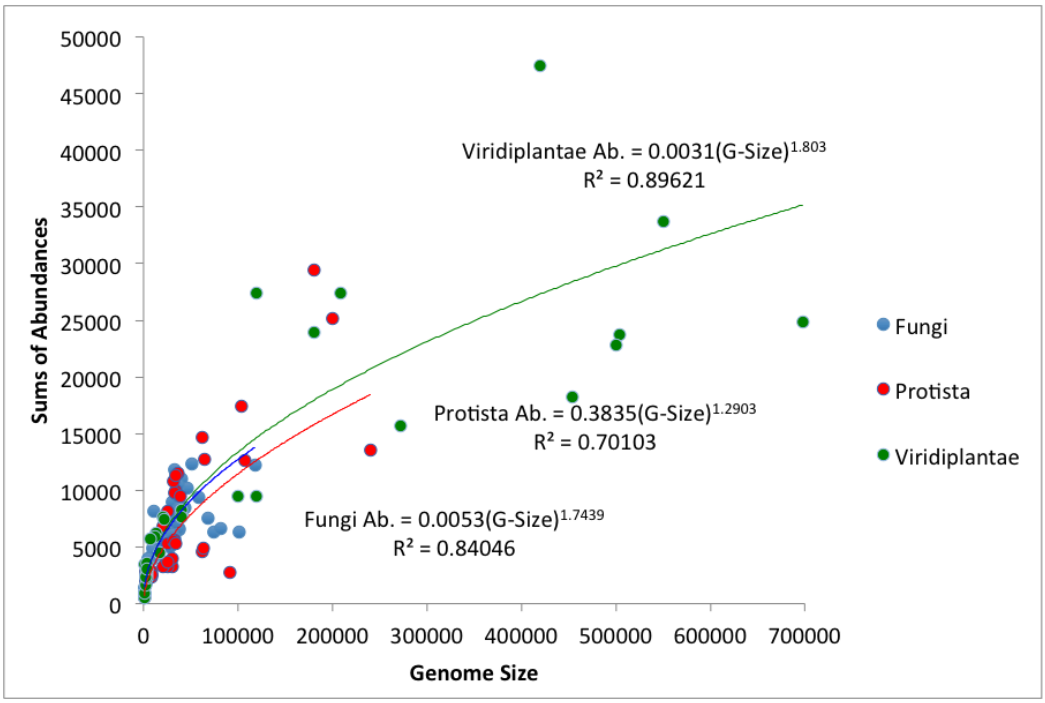

Figure 4.5: Eukaryotic Sums of Fold Abundances vs Genome Size for Fungi, Plants, and Protista

4.2 No Comb Bias

As discussed in section 3.4, justification for the lack of bias in the comb approach was sought by binning the count of nd for each of the comb nodes. A uniform distribution was discovered, with one spike near the combinatorial explosion, before it for FSFs and after it for folds, since a larger *nd* corresponds to an earlier time. The presence of this spike near the combinatorial explosion is discussed in section 4.3. However, the presence of a uniform distribution across the entire timeline suggests that the comb approach will not yield a biased perspective. The graph for both FSF and fold nodes is presented in Figure 4.6 below.

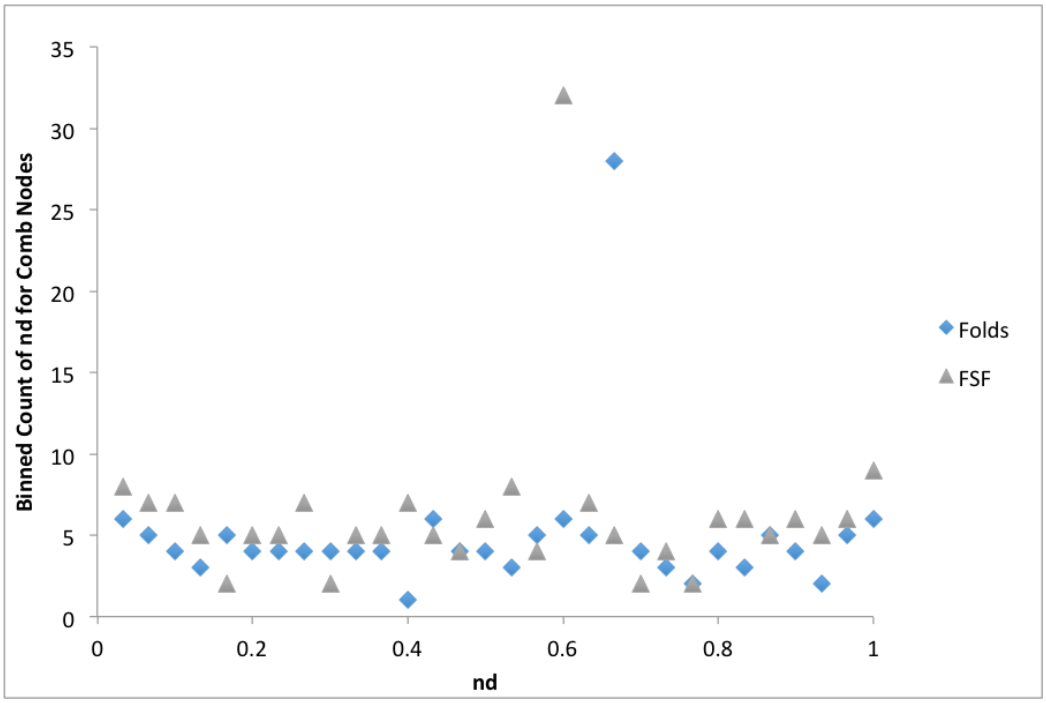

Figure 4.6: Binned count of normalized nd for *comb nodes* for FSF and fold data. Comb nodes are leaves diverging directly off the stem line on the tree of FSFs or folds. The bins are of size $\frac{1}{30}$.

4.3 Graphs for N_j, λ_j, and a_j and Their Interpretation

As described in chapter 3, both the a's and λ's were graphed as functions of the time of origin of a structure. The results for both FSFs and folds are presented in Figures 4.7-4.9.

4.3.1 N_j's

Figure 4.7 shows the $\log(N_j)$ vs time for both folds and FSFs for present day perfect comb nodes. Note that the present is at $t = 0$. Qualitatively, both fold and FSF total abundances have experienced a similar history, with total abundance dropping with time until a spike occurs around 1.5 Gyrs, at which point total abundance increases until the present. FSFs experienced this spike just prior to 1.5 Gyrs while folds experienced it just after. In general, also, FSFs had a larger total abundance over time.

Since only present day structures are used in this graph, the time on the x-axis represents a given structure's time of origin. Thus, in general, present

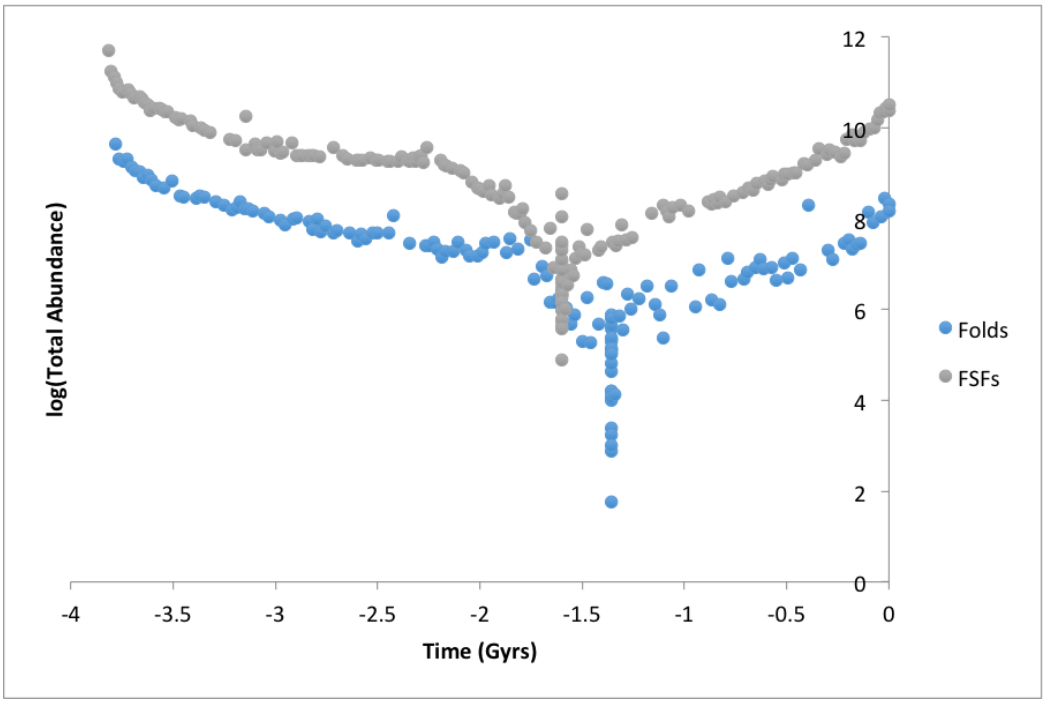

Figure 4.7: $\log(N_j)$ vs time (in Gyrs) for perfect comb fold and FSF nodes

structures that came into existence earlier have had longer to grow in abundance and, therefore, have a greater total abundance than structures that came into existence later. However, after the spike in the data innovations in the form of structure variations measured by λ's (see Figure 4.8) grow so rapidly that the newer structure abundances actually catch up to and outpace older structure abundances.

Moreover, the spike in the graph is present in all Figures, 4.7-4.9, always just prior to 1.5 Gyrs in FSFs and just after that time in folds. This time period corresponds to the combinatorial explosion of structural domains, or "big bang" discussed in [17], referring to the sudden, punctuated appearance of a large number of terminal leaves (structures) following the evolutionary halfway mark, primarily due to the high combination and rearrangement of domains and distribution of these new combinations in novel lineages, and the rise of organismal diversification.

4.3.2 λ_j's

Figure 4.8 shows the $\log(\lambda_j)$ vs time for both folds and FSFs for present day perfect p-comb nodes. Note again that the present is at $t = 0$. Again, qualitatively, both fold and FSF total abundances have experienced a similar history, with λ's increasing with time until a spike occurs around 1.5 Gyrs, at which point λ's increase much faster until the present. FSFs experienced this spike just prior to 1.5 Gyrs while folds experienced it just after. In general, also, FSFs had larger λ's over time. As previously mentioned, 1.5 Gyrs corresponds to the combinatorial explosion.

The super-exponential growth in λ's after the combinatorial explosion of structural domains is to be expected for the following reasons. Consider a particular structure X. Once structure combinations are possible X will appear in combination with other structures. Moreover, it is this new combination that gets selected for its functionality. That additional selection pressure, above that of X's on its own, speeds that structure's diversification. As combinations proceed and generations of new structures are born the new structures form combinations with each other as well as with structures prior to the explosion. Again, each new combination and structure that X plays a role in has a unique history and selection pressure that contributes to the λ explosion after the combinatorial explosion in Figures 4.8. This same super-exponential tendency in the λ's following the combinatorial explosion is also the reason why newer structure abundances outpace older structure abundances around this time in 4.7.

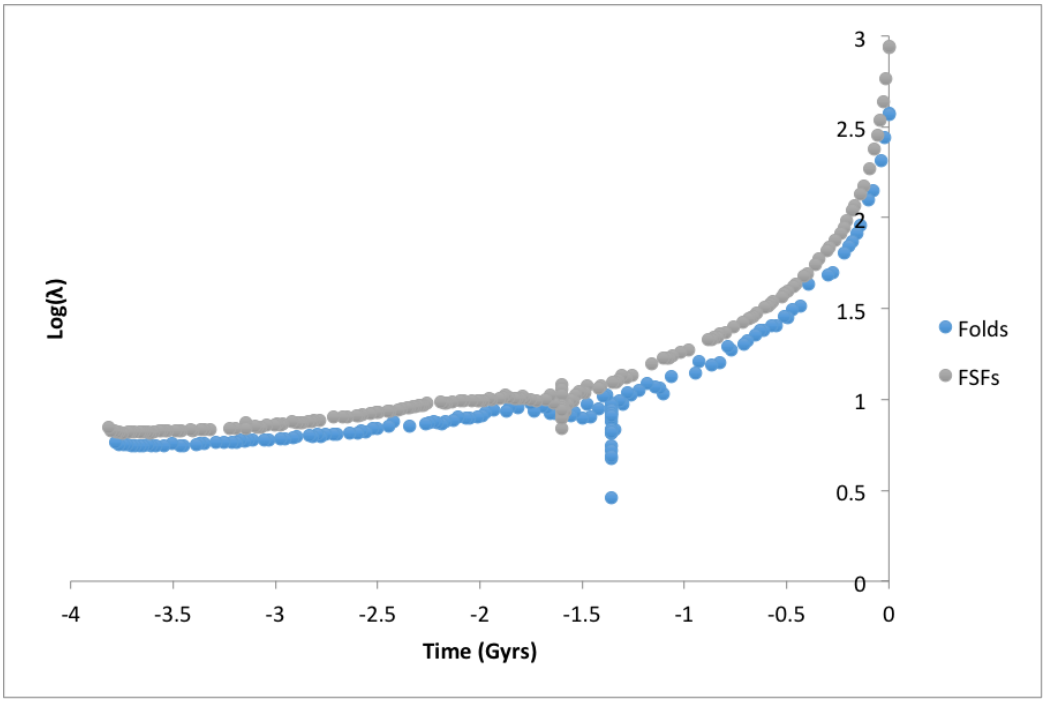

Figure 4.8: $\log(\lambda)$ vs time (in Gyrs) for perfect, p-comb fold and FSF nodes

4.3.3 a_j's

Figure 4.9 shows the $\log(a_{j'j})$ vs time for both folds and FSFs for present day perfect p-comb nodes. Note again that the present is at $t = 0$. Again, qualitatively, both fold and FSF total abundances have experienced a similar history, with a's increasing with time until a spike occurs around 1.5 Gyrs, at which point a's decrease until the present. FSFs experienced this spike just prior to 1.5 Gyrs while folds experienced it just after. Interestingly, unlike in Figures 4.7 and 4.8 where FSFs dominate folds, folds had larger $\ln(a_{j'j})$'s over time. This is because a's and N's have an inverse square relationship as seen in equation 2.14, and FSFs dominate folds in total abundance over time. Thus, folds dominate FSFs in $\log(a_{j'j})$'s over time because their total abundances are generally lower.

Domination aside, the values of a's increase consistently for both structures until the spike. This spike occurs nearer the end of superkingdom specification and into the epoch of organismal diversification [18]. The analogy we keep in mind is that of a developing field of knowledge. The initial researchers cannot help but make many novel discoveries, fundamental results and so forth, laying many flags in the sand. The next wave of researchers,

however, usually do not make as many wholly novel, fundamental discoveries, partly perhaps due to the scarcity of such results in the field, but also because it is easy to establish oneself by combining many of the basic results into novel combinations. Matters appear to have progressed in much the same way in the history of protein structure.

Thus, following the combinatorial explosion a's values decrease. As described earlier, old structures' appearances in various combinations results in high numbers of old structures being preserved relative to low numbers of entirely new structures. This explains the simultaneous rise in λ in Figure 4.8 and fall of a in Figure 4.9 following the combinatorial explosion. It also suggests that once combinations arose with a vengeance the easiest route to satisfying evolutionary needs was via combinations rather than novel domain discoveries. This makes a certain mathematical sense accurately captured in the phrase "combinatorial explosion" itself, combinations now yielding an avenue with a much larger number of possibilities than the alternative.

It should be clear that, qualitatively, λ's, a's, and N's for both FSFs and folds experienced very similar qualitative histories. "There is no doubt that protein families and superfamilies are monophyletic, that is, they derive from a common ancestor. In contrast, monophyly of protein folds, as opposed to folds originating by convergence from unrelated ancestors, remains an issue of debate" [19]. Our results suggest that folds may indeed retain an evolutionary relationship after all. This is not unexpected as a tight correlation between folds and FSFs has been previously noted in several studies [4],[17], and [18].

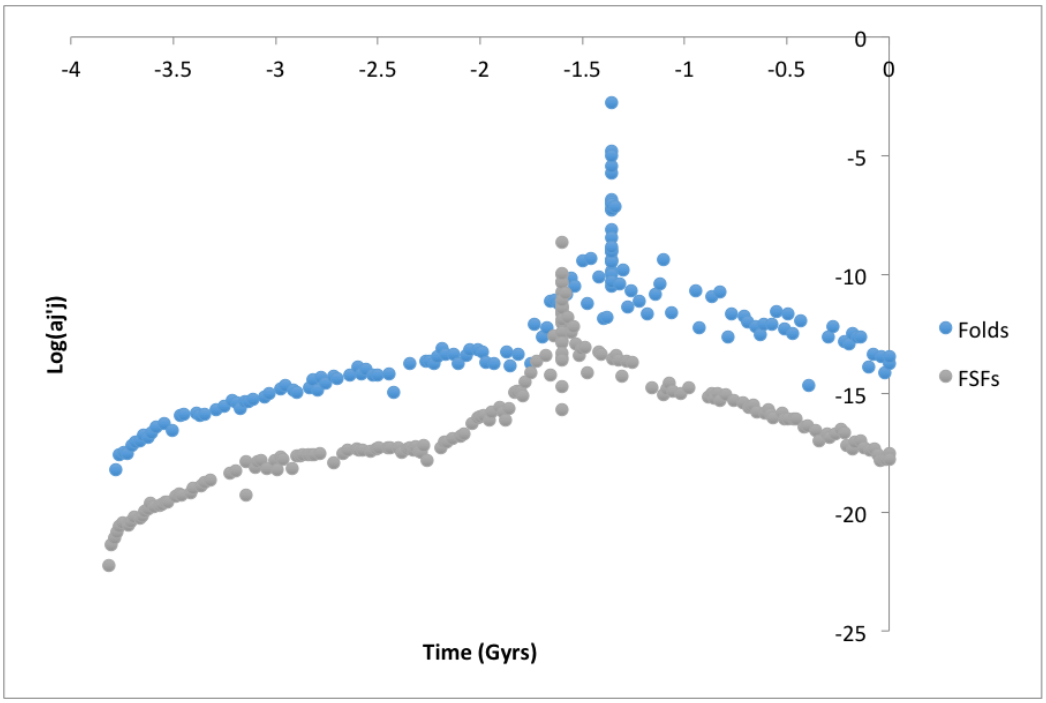

Figure 4.9: $\log(a_{j'j})$ for present j to future j' vs time (in Gyrs) for perfect, p-comb fold and FSF nodes

4.4 Comparing Folds to FSFs

Data from the SCOP 1.75 database reveals that there is, on average, between 1 to 2 FSFs in each fold, with an overall average of 1.64 FSFs per fold across all protein classes. Thus, it is no surprise that the history of folds and FSFs is quite similar in Figures 4.7-4.9. If there was only one FSF in each fold, then the λ of the FSF would be the same as the λ of the fold. This equality would then make the N's and a's similar by virtue of equations 2.7 and 2.11.

Yet, there is a gap between the two curves, especially evident in Figure 4.7 and Figure 4.9. Moreover, in all curves, the spike in the data occurs at slightly different times, just before 1.5 Gyrs in FSFs and just after that in folds. However, as mentioned in section 3.2, while the graphs of folds and FSFs are laid alongside each other they are not directly comparable. The significantly larger number of organisms in the FSF study, as compared to the fold study, accounts for the gap between the curves in Figure 4.7, causing the total abundance of FSFs, which is a function of the number of organisms, to be consistently above the total abundance of folds. The gap in the total abundance causes the gap in the remaining graphs due to the mathematical

relationships between N, λ, and a. Specifically, equation 2.11 shows a direct proportionality between $\log(N)$ and λ, resulting in FSFs being above folds in Figure 4.8. Since λ is only proportional to $\log(N)$ the visible gap shrinks. Similarly, equation 2.7 shows an inverse relationship between N and a so in Figure 4.9 we expect the fold curve to dominate the FSF curve.

Finally, note that the time in Figures 4.7-4.9 was determined via the linear relationship between t and nd in equations 2.6 and 2.7. However, nd is a *normalized* node distance so the larger the number of structures present in the study the larger the normalization. A larger normalization pushes back the nd value, and thus the time, of older structures. Since there are more FSFs than folds in our study, the FSFs have a larger normalization and this is a contributor to the FSF curve having a downward spike earlier than the fold curve in Figure 4.7. Again, the mathematical relationships between N, λ, and a then maintain this spike separation in Figures 4.8 and 4.9. These reservations on comparison aside, there is still a remarkable similarity between the evolutionary history of folds and FSFs evinced in the figures.

4.5 Conclusions

We have produced an approximate model for the evolution of protein folds and FSFs. We believe that the biological assumptions incorporated into our model are more plausible than those used in previous models. We used them to conclude: 1) that there appears to be a tight connection between the history of folds and FSFs, 2) that the corresponding transition probabilities to new variants of a fold experienced a sharp increase just as the transition probabilities to new folds experienced a steep decline and 3) that this simultaneous sharp increase in λ and decline in a is explainable by and consistent with the combinatorial explosion and the rise of organismal diversification. We believe that variations of our simple model will be applicable to other problems dealing with the evolution of protein structure, and may have uses wherever the simple assumptions of a birth-death model with available abundances and a molecular clock exist.

REFERENCES

[1] A. Nasir, A. Naeem, M. J. Khan, H. D. Lopez-Nicora, and G. Caetano-Anollés, "Annotation of protein domains reveals remarkable conservation in the functional make up of proteomes across superkingdoms," *Genes (Basel).*, vol. 2, no. 4, pp. 869–911, 2011.

[2] G. Caetano-Anollés, M. Wang, D. Caetano-Anollés, and J. E. Mittenthal, "The origin, evolution and structure of the protein world." *Biochem. J.*, vol. 417, no. 3, pp. 621–637, 2009.

[3] A. Nasir, "Origin Of Viruses Revealed By The Genomic Study Of Protein Domain Structures," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2012.

[4] D. Caetano-Anollés, K. M. Kim, J. E. Mittenthal, and G. Caetano-Anollés, "Proteome evolution and the metabolic origins of translation and cellular life," *J. Mol. Evol.*, vol. 72, pp. 14–33, 2011.

[5] J. Qian, N. M. Luscombe, and M. Gerstein, "Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model." *J. Mol. Biol.*, vol. 313, no. 4, pp. 673–81, Nov. 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11697896

[6] G. P. Karev, Y. I. Wolf, A. Y. Rzhetsky, F. S. Berezovskaya, and E. V. Koonin, "Birth and death of protein domains : A simple model of evolution explains power law behavior," *BMC Evol. Biol.*, vol. 26, 2002. [Online]. Available: http://www.biomedcentral.com/1471-2148/2/18

[7] A. Magner, W. Szpankowski, and D. Kihara, "On the Origin of Protein Superfamilies and Superfolds," *Sci. Rep.*, vol. 5, p. 8166, 2015. [Online]. Available: http://www.nature.com/doifinder/10.1038/srep08166

[8] K. B. Zeldovich, I. N. Berezovsky, and E. I. Shakhnovich, "Physical origins of protein superfamilies," *J. Mol. Biol.*, vol. 357, no. 4, pp. 1335–1343, 2006.

[9] K. B. Zeldovich, P. Chen, B. E. Shakhnovich, and E. I. Shakhnovich, "A first-principles model of early evolution: Emergence of gene families, species, and preferred protein folds," *PLoS Comput. Biol.*, vol. 3, no. 7, pp. 1224–1238, 2007.

[10] G. Caetano-Anollés and D. Caetano-Anollés, "An evolutionarily structured universe of protein architecture an evolutionarily structured universe of protein architecture," *Genome Res.*, pp. 1563–1571, 2003.

[11] G. Caetano-Anollés and D. Caetano-Anollés, "Universal sharing patterns in proteomes and evolution of protein fold architecture and life," *J. Mol. Evol.*, vol. 60, no. 4, pp. 484–498, 2005.

[12] A.-R. Carvunis, T. Rolland, I. Wapinski, M. a. Calderwood, M. a. Yildirim, N. Simonis, B. Charloteaux, C. a. Hidalgo, J. Barbette, B. Santhanam, G. a. Brar, J. S. Weissman, A. Regev, N. Thierry-Mieg, M. E. Cusick, and M. Vidal, "Proto-genes and de novo gene birth," *Nature*, pp. 3–8, 2012.

[13] A. Caballero, "Developments in the prediction of effective population size." *Heredity (Edinb).*, vol. 73 (Pt 6), no. March, pp. 657–679, 1994.

[14] M. Lynch and J. S. Conery, "The origins of genome complexity." *Science*, vol. 302, no. 5649, pp. 1401–1404, 2003.

[15] M. Lynch, "The origins of eukaryotic gene structure," *Mol. Biol. Evol.*, vol. 23, no. 2, pp. 450–468, 2006.

[16] M. Wang, Y. Y. Jiang, K. M. Kim, G. Qu, H. F. Ji, J. E. Mittenthal, H. Y. Zhang, and G. Caetano-Anollés, "A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation," *Mol. Biol. Evol.*, vol. 28, no. 1, pp. 567–582, 2011.

[17] M. Wang and G. Caetano-Anollés, "The Evolutionary Mechanics of Domain Organization in Proteomes and the Rise of Modularity in the Protein World," *Structure*, vol. 17, pp. 66–78, 2009.

[18] M. Wang, L. S. Yafremava, D. Caetano-Anollés, J. E. Mittenthal, and G. Caetano-Anollés, "Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world," *Genome Res.*, vol. 17, no. 11, pp. 1572–1585, 2007.

[19] E. V. Koonin, Y. I. Wolf, and G. P. Karev, "The Structure of the Protein Universe and Genome Evolution," vol. 420, no. November, 2002.