

© 2022 Muhammad Fayeز Aziz

TRACING HISTORY OF PROTEIN DOMAIN ORGANIZATION, FUNCTIONAL LOOPS,
AND DIPEPTIDES WITH CHRONOLOGIES AND NETWORKS

BY

MUHAMMAD FAYEZ AZIZ

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Crop Sciences
in the Graduate College of the
University of Illinois Urbana-Champaign, 2022

Urbana, Illinois

Doctoral Committee:

Professor Gustavo Caetano-Anollés, Chair
Professor Sandra Rodriguez-Zas
Professor Maria Villamil
Adj. Associate Professor Kaustubh Bhalerao

ABSTRACT¹

Systems biology is a multidisciplinary, holistic, and rising field that uses computational and mathematical approaches to model and explore the makeup and functioning of biological systems. Analysis of high throughput data with statistical approaches help create in silico models that advance our understanding of complex biological phenomena, with illuminating findings in human development and disease. Also, drawing links between biochemistry and physiology determine how functions take place in biological interactions. We here focus on the mysterious origin and evolution of proteins in their various forms. A hierarchical classification of proteins based on homology of structural domains describes the extant shared-and-derived features of the sequence, structure, and function of protein modules. Modern advances in evolutionary genomics and systems biology enable the historical exploration of the structure, function, and organization of proteins. The history of a biological system can be traced through a series of time events, which define molecular chronologies (timelines) describing the rise of biological innovations. Overlaying these histories on time-varying networks provide a window into the past. Phylogenomic analyses allow construction of molecular timelines of evolving networks, which can be linked to the geological fossil record using a molecular clock of folds. These evolutionary timelines of domains ('lock-and-key' folding components of proteins) reveal the explosive emergence of multidomain proteins, yin-and-yang patterns of 'elementary functionome' of loops

¹ Materials in this document are reprinted, with permission, from Caetano-Anollés G, Aziz MF, Mughal F and Caetano-Anollés D, 2021, "Tracing protein and proteome history with chronologies and networks: folding recapitulates evolution (Abstract, Introduction)," *manuscript published*, Caetano-Anollés G, Aziz MF, Mughal F, Gräter F, Koç I, Caetano-Anollés K and Caetano-Anollés D 2019, "Emergence of Hierarchical Modularity in Evolving Networks Uncovered by Phylogenomic Analysis (Abstract, Introduction)," *manuscript published*, Aziz MF and Caetano-Anollés G, 2021, "Evolution of Networks of Protein Domain Organization (Chapter 2)," *manuscript published*, and Aziz MF, Mughal F and Caetano-Anollés G, "Untangling the birth of structural domains by recruitment of protein loops with chronologies, networks, and *ab initio* structural prediction (Chapter 3)," *manuscript in preparation*

and domains (threads uniting domains) development, and a late rise of translational proteins and dipeptides (pairs of amino acids making up domains). Growing networks of these primordial biological entities describe how parts associate with each other to form integrated systems, which are often structured by modularity, power law or a modular hierarchy (i.e., both). A ~2,000-year-old papyrus attributed to *Empedocles* from the ancient city of Panopolis in Upper Egypt, *P. Strasb. Gr. Inv.* 1665-6, recounts a ‘double tale’ of unification and change. This scribal transmission can astonishingly be interpreted as a description of biological evolution with network hierarchies ~2,400 years before Darwin. We discover that the double tale portrays a biphasic (hourglass) theory of module emergence. A first phase links parts weakly, which then associate variously. Later, these parts diversify and compete, and are often selected for performance. The emerging interactions constrain their associative structure, which causes parts to self-organize into modules with tight linkages. In a second phase, variants of the modules diversify and become parts for a new generative cycle of higher-level organization. This paradigm predicts the rise of hierarchical modularity in evolving networks at different timescales and complexity levels. Remarkably, phylogenomic analyses uncover this emergence in the evolving networks of protein domain organization, elementary functionome and dipeptides. We use computational techniques to mine historical developments and observe how 1,937 functional loop prototypes and 400 structural elements of dipeptides define function and modular behavior of 1,442, 1,475 and 6,162 protein domains. Evolutionary developments of protein domains describe macro level structural evolution, as protein sizes may be different among species, but their active sites can be similar. The crystal structures of protein architectures are highly conserved and thus the patterns of historical interactions among these signified the co-option and recruitment of functional modules throughout protein history. Timelines built using modern methods of phylogeny reconstruction

and mapped over evolving networks of domain structures and architectures described the early emergence of domains involving sandwich, bundle, barrel, and other complex structures, and a late ‘big bang’ of domain combinations through evolutionary processes of fusion and fission. Visualizations as radial and waterfall layouts of timed networks of 6,162 protein domain and multi-domain superfamilies identified the development of domain clusters and significant hubs involved in determining affinity relationships of protein domains in the genomes. The network analyses made evident the scale free and modular behaviors of protein domain organization as proteins evolve through unique events in molecular history. Similarly, the bipartite and waterfall visualizations of the bipartite and projected networks of 1,937 loop prototypes and associated protein fold families described the sharing topology and evolution of an elementary functionome. We investigated how loops impart a particular function to domains and the diversity of domain function is represented by the variety of loops contained therein and shared with other domain structures. The bimodal networks and their monomodal network projections helped identify the hierarchical modularity of loop-domain, domain-loop, loop-loop, and domain-domain interactions that might be very ancient in nature and have been conserved throughout modern evolution. These networks hence helped to elicit the origin and evolution of structural domains and modular and non-modular loops through ‘waves’ of recruitment phases, promising strong genomic and structural evidence, with a potential to model the gradual evolution of proteins, as we illustrated with AlphFold2-modelled P-loop hydrolases. Lastly, we study the 400 dipeptides and associated protein domain classes and genomes to trace how evolution of protein structure combines supersecondary structures and dipeptides into domains, developing complex translation machinery and enhancing structural complexity by establishing long-distance interactions in novel structural and architectural designs. The dipeptides define functional aspects of cells from translational

machinery to protein binding and thus can play an extremely significant role in determining the evolutionary history of primordial protein structures and associated processes. The bipartite networks of domains and genomes with dipeptides, and the projected networks of dipeptides therein, portray the long standing and conserved relationships among these symbiotic entities. Further investigation highlights structural features of evolutionary significance and describes how the structural constitution of domains have been defined by these dipeptides of archaic nature and have transformed the functional landscape of protein expression.

To Asma, Rania, Rayan and Amaan

ACKNOWLEDGMENTS

This project has been possible with utmost support of my adviser, Prof. Dr. Gustavo Caetano-Anollés and spur from his caring wife Gloria. I thank Prof. Gustavo for his revisions, feedback, inspiration, confidence, and guidance along every step of the way. He allowed and enabled me to take up the challenge of doing a full-time job and work on my thesis in parallel. Speaking of the challenges during all this turmoil and the pandemic, no words can describe the diligent support from my wife Asma, with all her love and beauty, her sweating efforts to let me stay focused on the job, immigration and thesis while she took care of everything else from groceries to teaching our kids, and her sacrifices of leisure time and lack of attention from me, rather having to often console me when I touched the boundaries of sanity. A special honorable mention of my brilliant and delightful daughter Rania whom I owe a lot of missed playtimes together, and “who is as old as my thesis”, if I may quote my wife. To Rayan and Amaan, my sons, for cheering me up and to their youth which I look forward to making up for in my ‘post-Doc’ life. I respectfully recognize my committee members, Sandra Rodriguez-Zas, Maria Villamil, and Kaustubh Bhalerao, for discipline, cheer, and vision. Also, I acknowledge Dr. Minglei Wang, for the base data set used to generate the timelines, and Fizza Mughal, for pioneering the generation of a primary data set and curating analysis. I would also like to acknowledge the funding support received from the National Science Foundation and the United States Department of Agriculture, and initial support from COMSATS University, Pakistan. Thanks to my departed father for imparting his valuable knowledge and polishing my skills, my mother for the foundational nourishment and prayers, and my two older PhD brothers for the inspiration, suggestions, and emotional support. I thank my family friends for reminding me the value of this and keeping me going. Finally, all praise and thanks to Allah (God), the Almighty, the Graceful, the Kind, for bringing about everything.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION	1
1.1 Protein Structure	1
1.2 Protein Domains and Protein Classification	2
1.3 Comparison of Structural Classification Systems	3
1.4 Protein Domain Organization.....	3
1.5 Evolution of Domains and Domain Organization	4
1.6 The Evolutionary Timeline of Domain Organization and the Ages of Domains and Domain Combinations.....	6
1.7 Rationale behind Investigating Networks of Protein Domain Organization	9
1.8 Exploration of Biological Networks to Elaborate the Role of Protein Domains in Evolution	11
1.9 Organization of Proteins by Loop and Dipeptide Subunits	21
1.10 Scale-free and Sister Explanations of the Growth Patterns of Biological Networks.....	25
1.11 The Focus of this Study	33
1.12 Introductory Figures.....	36
1.13 Introductory Tables.....	47
1.14 Introductory References	47
CHAPTER 2: EVOLUTION OF NETWORKS OF PROTEIN DOMAIN ORGANIZATION	58
2.1 Abstract.....	58
2.2 Introduction.....	59
2.3 Results and Discussion.....	64
2.4 Conclusions.....	77
2.5 Methods	78
2.6 Figures	86
2.7 Tables.....	93
2.8 References.....	95
CHAPTER 3: UNTANGLING THE BIRTH OF STRUCTURAL DOMAINS BY RECRUITMENT OF PROTEIN LOOPS WITH CHRONOLOGIES, NETWORKS, AND <i>AB INITIO</i> STRUCTURAL PREDICTION.....	102
3.1 Abstract.....	102
3.2 Introduction.....	103
3.3 Results and Discussion.....	107
3.4 Conclusions.....	132

3.5 Methods	134
3.6 Figures	141
3.7 References.....	150
CHAPTER 4: A NETWORK VIEW OF DIPEPTIDE EVOLUTION	157
4.1 Abstract.....	157
4.2 Introduction.....	158
4.3 Results and Discussion.....	160
4.4 Conclusions and Future Directions	181
4.5 Methods	181
4.6 Figures	193
4.7 References.....	160
CHAPTER 5: CONCLUSIONS.....	208
APPENDIX A: LIST OF ACRONYMS	214
APPENDIX B: SUPPLEMENTARY INFORMATION FOR CHAPTER 2.....	216
B.1 Supplementary Text	216
B.2 Supplementary Materials and Methods	216
B.3 Supplementary Figures.....	216
B.4 Supplementary Tables	216
B.5 Supplementary Videos	216
B.6 Supplementary Data Files.....	216
APPENDIX C: SUPPLEMENTARY INFORMATION FOR CHAPTER 3	217
C.1 Supplementary Materials and Methods	217
C.2 Supplementary Figures.....	217
C.3 Supplementary Videos	217
APPENDIX D: SUPPLEMENTARY INFORMATION FOR CHAPTER 4	218
D.1 Supplementary Materials and Methods	218
D.2 Supplementary Figures.....	218
D.3 Supplementary Videos	218

CHAPTER 1: INTRODUCTION

“... I saw the Aleph from every vantage point, I saw in the Aleph the earth and in the earth again the Aleph, I saw my face and viscera, I saw your face and in vertigo I wept, for my eyes had seen that secret and conjectural object whose name is usurped by men ...”— Jorge Luis Borges, *The Aleph and Other Stories*

1.1 Protein Structure

Proteins are relatively long polypeptides that fold into globular or fibrous forms. A polypeptide is a single linear polymer chain of amino acids joined together by peptide bonds that covalently link the carboxyl and amino groups of adjacent amino acid residues. Proteins facilitate biological functions sometimes by associating together to form stable protein complexes. The foundation of these associations is the folding of the linear polypeptide chain into regular three-dimensional (3D) structural arrangements of the amino acid backbone. These arrangements form helices (mostly alpha helices), beta sheets, turns and bends that are occasionally separated by non-regular loop segments. Secondary structure elements of these kinds were already described in 1951 by Linus Pauling and colleagues (Pauling, Corey and Branson, 1951). They are mostly defined by patterns of hydrogen bonds between the main-chain peptide groups when made explicit by X-ray crystallography and NMR spectroscopy (Murzin *et al.*, 1995). Peptide groups of size 2, i.e. dipeptides, have been argued to be arranged as non-regular and flexible loop structures to impart molecular functions and genetics (Caetano-Anollés, Wang and Caetano-Anollés, 2013). Striking regularities have been shown to exist in how secondary structures are assembled (Levitt, Chothia and others, 1976; Chothia, Levitt and Richardson, 1977) and in the topologies of the polypeptide

chains (Richardson, 1976, 1977; Sternberg et al., 1976). These regularities result from the intrinsic physical and chemical properties of proteins (Chothia, 1984; Finkelstein and Ptitsyn, 1987).

1.2 Protein Domains and Protein Classification

Protein domains have been regarded as canonically conserved and fundamental units of protein classification (Murzin *et al.*, 1995). As of February 24, 2018, there were 137,917 domains among the 128,043 protein structures in the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB-PDB). Domains are structural and functional units of evolution (Apic, Gough and Teichmann, 2001b) and one or more domains make up proteins, sometimes in unusually complex arrangements (Wang and Caetano-Anollés, 2009). Domains fold into compact 3D structures that arrange the alpha-helical and beta-sheet secondary structure elements into tightly packed conformations of the polypeptide chain. Since proteins often contain more than one domain, domains appear repeatedly in the protein molecules, singly or in combination with other domains (Wang and Caetano-Anollés, 2009). Domains are not only recurrent but are also highly conserved (Bajaj and Blundell, 1984). They are regarded as bona fide structural, functional and evolutionary units in structural classification schemes of proteins (Murzin *et al.*, 1995). Structural domains have been widely used in the study of evolution, including the analyses of biological network evolution, because protein structure exhibits clearer domain definitions and evolutionary family relationships than genome sequences (Apic, Gough and Teichmann, 2001b). The Structural Classification of Proteins (SCOP) is a popular taxonomy gold standard (Murzin *et al.*, 1995) that provides superfamily definitions of domains, which can be used to scan genome sequences for motifs of domains in terms of their superfamily combinations (Apic, Gough and Teichmann, 2001b). Only a limited repertoire of domain superfamilies is available (Chothia and others, 1992;

Wolf, Grishin and Koonin, 2000). In SCOP, the structure of protein domains exhibiting similar 3D arrangements of secondary structures and thus identical topologies have been classified as folds (F) (Murzin *et al.*, 1995). Within folds, protein domains whose structure and functional features indicate a common evolutionary origin are further grouped into fold superfamilies (FSF). For example, the ATPase domain of the heat-shock protein is an FSF that comprises hexokinases and actins (Flaherty *et al.*, 1991). Finally, FSFs with more than 30% amino acid residue similarity are further classified into fold families (domain) (Fig. 1.1).

1.3 Comparison of Structural Classification Systems

The conservation of molecular features in FSF domains throughout billions of years of evolution highlights the significance of the FSF classification level. Even with the growing discovery of domains, the number of Fs and FSFs remains virtually the same. This trend suggests that as evolution proceeds, the discovery of new building blocks is limited and biased towards details of the fold structure, including the decoration of main 3D fold designs. A comparison of SCOP with other protein structure classification schemes helps illustrate this realization (Table 1.1). There is yet a need to understand the structural complexities of the protein world from an evolutionary perspective. One approach is to retrodict the emergence of modern protein domain structure from earlier structural states using phylogenetic principles. This must fulfill the evolutionary principle of spatiotemporal continuity (*'lex continui'* of Leibnitz) (Nasir, Kim and Caetano-Anollés, 2017), a fact that will become evident later on in our discussions.

1.4 Protein Domain Organization

The evolution of protein structure has been influenced significantly by how domains 'organize'

along the amino acid sequence chain. More than two-thirds of the protein sequences are longer than an average domain length, which implies that a vast majority of genomic proteins are multidomain proteins (M Gerstein, 1998) that have evolved through gene duplication, recombination and fusion (Apic, Gough and Teichmann, 2001b). The arrangement of domains along the sequence of multimeric proteins is referred to as '*domain organization*'. The atomic structure of domains along with domain organization have been collectively termed '*domain architecture*' (Fig. 1.2). Both the structure and organization of domains are considered far more conserved than protein sequence (Wang and Caetano-Anollés, 2009). For instance, tyrosine kinases exhibit highly conserved domain arrangements at the subfamily level as they distribute along metazoan lineages (Shiu *et al.*, 2004).

1.5 Evolution of Domains and Domain Organization

Protein domains are conserved evolutionary units organizing as complex arrangements in various but a finite number of combinations (Bashton and Chothia, 2002; Wang and Caetano-Anollés, 2006a). The domain combinations constitute functional modules that recur in different protein contexts and have been termed supradomains (Vogel, Berzuini, *et al.*, 2004). Here, modules have parts that interact with each other more than with other parts or modules of the system. Evolutionary studies of domain organization include analyses of conservation and variation of domain associations (Bashton and Chothia, 2002; Apic, Huber and Teichmann, 2003; Vogel, Bashton, *et al.*, 2004; Vogel, Berzuini, *et al.*, 2004; C Vogel, Teichmann and Pereira-Leal, 2005), mechanisms of generation of new domain combinations (Enright *et al.*, 1999; Marcotte *et al.*, 1999; Bornberg-Bauer *et al.*, 2005; Bashton and Chothia, 2007), difference between fusion and fission mechanisms (Enright *et al.*, 1999; Marcotte *et al.*, 1999; Yanai, Derti and DeLisi, 2001;

Moore and Bornberg-Bauer, 2012; Bornberg-Bauer and Albà, 2013; Moore *et al.*, 2013), circular permutations in multidomain proteins (Jeltsch, 1999; Moore *et al.*, 2008; Chothia and Gough, 2009; Bornberg-Bauer, Huylmans and Sikosek, 2010; Liberles *et al.*, 2012), and domain insertion and loss (Aroul-Selvam, Hubbard and Sasidharan, 2004; Liberles *et al.*, 2012; Bornberg-Bauer and Albà, 2013). Most of these studies have adopted statistical or mechanistic approaches. Very few have embraced global phylogenomic views. One of the challenges is to find a general metric of pairwise comparison because of problems in the systematic classification of structural topology, even when the domain structures enable to extract evolutionary information (W R Taylor, 2007).

One approach to investigate protein evolution takes advantage of the reconstruction of phylogenomic trees of single domain and multidomain proteins based on the occurrence and abundance of domain structures in proteomes at F and FSF levels (Caetano-Anollés and Caetano-Anollés, 2003; Wang and Caetano-Anollés, 2006a; Wang *et al.*, 2007) (Fig. 1.3). These rooted global trees have architectural lineages as branches leading to structural leaves. They have been utilized to discover and analyze the evolutionary history of proteins (Caetano-Anollés and Caetano-Anollés, 2003, 2005; Wang and Caetano-Anollés, 2006a), origins and evolution of metabolic networks (Caetano-Anollés, Kim and Mitterenthal, 2007), and reductive tendencies in architectural repertoires linked to origins of diversified life (Wang *et al.*, 2007). Having been based on a genomic census of domain structure (Caetano-Anollés and Caetano-Anollés, 2003; Yang, Doolittle and Bourne, 2005; Wang *et al.*, 2007) and organization (Wang and Caetano-Anollés, 2006a; Fukami-Kobayashi *et al.*, 2007), these phylogenomic trees not only summarize the history of combinatorics of domain organization through fusion and fission events that define the evolutionary mechanics of protein domain architecture (Wang and Caetano-Anollés, 2009), but also reveal the tripartite nature of the living world (Wang and Caetano-Anollés, 2006a). Moreover,

the results uphold the nonrandom nature of domain combinatorics being shaped by natural selection under an optimality criterion (Wang and Caetano-Anollés, 2006b). Correspondingly, the domain fusion and fission processes surrounding multimeric proteins proceed at different relative rates (Kummerfeld and Teichmann, 2005; Pasek, Risler and Brézellec, 2006). It is therefore important to trace the resulting architectures directly in the phylogenomic trees to eventually uncover evolutionary patterns related to a multiplicity of biological functions, especially those associated with the explosive expansion of the repertoire of domain combinations occurring relatively late in evolution (Wang and Caetano-Anollés, 2009).

1.6 The Evolutionary Timeline of Domain Organization and the Ages of Domains and Domain Combinations

The trees of domain architectures reconstructed from the global architectural census embody timelines (chronologies) of appearance of architectures during protein evolution and thus reveal fundamental patterns in the history of the protein world. These chronologies of molecular discovery are influenced by the dominance of domain fusion and fission processes throughout evolution (Wang and Caetano-Anollés, 2009). The 3D architectural designs, which remain static over extended periods of time (Chothia *et al.*, 2003) harbor deep phylogenomic signals (Caetano-Anollés and Caetano-Anollés, 2003). We previously reconstructed evolutionary timelines of molecular accretion using a phylogenomic framework, where molecular makeup was defined by accumulation of substructural or modular parts. Fig. 1.4 illustrates the universality of the accretion process, which extends not only to molecular components such as the synthase complex but also to products of mankind such as cities and high-rise buildings. Phylogenomic studies focus on the formation of domain combinations during evolution and describe the global history of a molecular

interactome based on abundance of architectures in hundreds of proteomes. These architectures have been mainly categorized as domains (singly occurring domains), domain pairs (two different domains), multidomains (numerous domains, with occasional repetition) and domain repeats (domains of one type that are repeated) (Wang and Caetano-Anollés, 2009). This diversification in protein architectures evolved through gradual accumulation, starting with a few architectures earlier in the timeline followed by an increasing rate of their proliferation. This trend culminated in an abrupt appearance of many terminal leaves with short branch lengths in the phylogenomic tree of domains and multidomains, an incident coined as the “big bang” of domain combinations after an analogy to a popular hypothesis for the cosmological origin of the universe. These timelines were generated based on the sequence and structure of thousands of nucleic acid molecules and millions of protein sequences encoded in hundreds of genomes. Remarkably, the architectural accumulation continues to date but with a decreasing rate of discovery of new architectures (Wang and Caetano-Anollés, 2009; Aziz, Caetano-Anollés and Caetano-Anollés, 2016).

The massive combinatorial exchange of modules that occurred late in evolution resulted in an enormous molecular and organismal diversity (Wang and Caetano-Anollés, 2009), suggesting the existence of an evolutionary transition (Maynard-Smith and Szathmary, 1995). The combinatorial mechanisms that yielded domain combinations or multidomains during the ‘big bang’ transition, primarily either recruited neighboring domains (fusional combinations), embedded neighboring sequences to release new evolutionary units (fusional/fissional combinations), or excised gene segments during chromosomal rearrangement to give rise to new modules (fissional combinations) (Wang and Caetano-Anollés, 2009) (Fig. 1.5). In a similar fashion, single domains involved in such a combinatorial interplay either fused with other single or multidomains to form combinations

(fusional domains), emerged through fission of combinations (fissional domains), originated from fission but engaged later in fusions (fissional/fusional domains), or occasionally remained inert during the combination processes (noncombinable domains). These mechanisms suggest a historical order of appearances of multidomains vis-à-vis the discovery of their domain constituents.

The phylogenomic tree of proteins reconstructed from the abundance of both domains and their combinations in proteomes embeds molecular chronologies (i.e. molecular history) and hence provides a basis for an evolutionary timeline of protein domain organization (Fig. 1.6). Since a fully unbalanced and rooted phylogenomic tree attributes each terminal leaf with its graphical distance from the root, these sequential numerical values determine the relative chronological order of the appearance of the leaves (i.e. the domains and multidomains). This property of trees allows formulation of a tangible timeline of protein domain evolution in terms of successive normalized node distance (*nd*) values (Wang and Caetano-Anollés, 2009) that represent the age of every domain and domain combination along the evolutionary timeline (Fig. 1.7). Use of a reference timeline like this in the present analysis of evolution of protein domains elucidated the gradual evolutionary appearance of protein domain structures (Caetano-Anollés *et al.*, 2009; Kim and Caetano-Anollés, 2010; Caetano-Anollés, Kim and Caetano-Anollés, 2012; Edwards, Abeln and Deane, 2013) and their combinatorial rearrangement in proteins (Wang and Caetano-Anollés, 2009). Chemical and biophysical properties of proteins have also been traced along the timeline. Examples include uncovering the natural history of biocatalysis by tracing chemical mechanisms in enzymatic reactions (Nath, Mitchell and Caetano-Anollés, 2014), evolutionary analysis of optimization and perpetual increase of protein folding speed through tracing a flexibility correlated factor known as contact order (the average relative distance of amino acid contacts in the tertiary

structure of proteins) (Debès *et al.*, 2013), and periodic induction of stochastic biphasic-rewiring and modularity in morphing metabolomics networks of *Escherichia coli*, within minutes after subjection to stress (Aziz *et al.*, 2012). Also included in this study was a reference timeline spanning billions of years (Gy) that was calibrated by a molecular clock of protein structures (Wang *et al.*, 2011). This timeline encompassed the origin of the protein world billions of years ago (Gya) and the present, thus serving as a guide to trace the evolutionary progression of domain networks in geological time.

1.7 Rationale behind Investigating Networks of Protein Domain Organization

Domains rarely operate in isolation making their interaction inevitable for vital organismal functions. Amino acid sequences, the building blocks of protein domains, of eight completely sequenced unicellular organisms have been sampled from the three superkingdoms of life (Archaea, Bacteria and Eukarya) and their sequence length distributions were observed (M Gerstein, 1998). More than two-thirds of the proteins were found to be longer than an average domain length, which implied that those proteins were multi-domain (M Gerstein, 1998; Wang and Caetano-Anollés, 2009). A similar study revealed that nearly two-thirds of the protein sequences from the bacterium *Mycoplasma genitalium* comprise of more than one domain (Teichmann, Park and Chothia, 1998). A census also showed that only one-fifth to one-third of the genomic proteins are single domain (Apic, Gough and Teichmann, 2001b). Thus, a majority of the proteins encoded in a genome are multi-domain, evolved through gene duplication, recombination and fusion.

In a similar analysis on 859 SCOP domains, 783 FSFs that were reported to occur in 1,307 pairwise combinations revealed a scale-free pattern (Teichmann, Chothia and Gerstein, 1999). Moreover,

most FSFs paired in only a few combinations and those with high genomic abundance were found in many combinations. P-loop nucleotide triphosphate hydrolases (c.37.1) and Rossmann domains (c.2.1) were the most versatile FSFs. Each of them showed a unique unshared repertoire of combination partners. Given current knowledge of domain structures and their evolutionary relationships, another study investigated the patterns of domain organization using a graph theoretical approach to reveal the evolutionary mechanisms behind the creation of complex proteins (Apic, Gough and Teichmann, 2001b). The network yielded by this unconventional method comprised of domains spliced from domain combinations and the compositional sharing relationship of the domains. The authors chose a network generation criterion in which two domains of a domain combination were connected with an edge (link or connection) if their mutual distance in the polypeptide chain was less than or equal to 30 residues. An interesting aspect of this criterion was the preservation of N to C-terminal orientation of the domain combination, which proclaims that during the process of recombination, the expensive alternative to evolve an interface and associated function twice was avoided in favor of duplication and then mutation of the domain combination in its original orientation (Apic, Gough and Teichmann, 2001b). Domain interactions in structural genomics were also included, keeping pace with the increasing library of PDB structures (Brenner and others, 2000; Burley and others, 2000), with the objective of gaining important insights into protein function and protein-protein interactions (PPIs) (Apic, Gough and Teichmann, 2001b).

In order to understand how a genome transitions into a set of proteins that is canalized (preserved) in evolution, the conserved repertoire of domains that combine and sometimes repeat in various ways must be analyzed. One focal aspect of our research is to discover domain interactions in molecular evolution and examine how these interactions change in time. Here, we derive the

protein domain interactome from the phylogenomic tree of domains and domain combinations (Wang and Caetano-Anollés, 2009) (Fig. 1.3). We also visualize the interactions using a graph theoretical approach (Apic, Gough and Teichmann, 2001b) to capture their historical development. Carefully selected criteria of connectivity among protein domains ensured a test framework with congruent evolutionary inferences from the resulting networks. One of the challenges was to achieve fruitful visualizations of the networks of domain organization that can uncover patterns of progression of connectivity along the evolutionary timeline. Various network layouts including radial, bipartite and waterfall, make evident the time flow of connectivity and also the clustering patterns of the nodes of the growing networks (Fig. 1.8). The network connectivity distributions were then inspected to help identify hubs of molecular combinatorial activity in the evolving domain networks. Domain structures with high connectivity scores in the distributions revealed the *donors* and *acceptors* of the recruitment processes that drive the evolution of protein domains. However, the most notable potential for a prime level investigation of evolution of domain organization was the exploration of network structure, the morphology of the preferential attachment principle (Barabási and Albert, 1999; Tanaka, 2005), and modular organization of evolving networks.

1.8 Exploration of Biological Networks to Elaborate the Role of Protein Domains in Evolution

Evolving networked systems are prevalent in biology and have recently gathered interdisciplinary interest (Watts, 2004). Networks have been used as an exploratory tool for the space of possible, meaningful and desirous linkages between parts of natural or engineered biological systems such as metabolism or PPI repertoires. The goal is to find interesting patterns that can help explain

processes operating in investigative studies about genetics, physiology and pathology. Complex networks that explain the internet or biological interactions occurring inside a cell, have been reviewed with a focus on statistical mechanics of network topology and dynamics, and a special consideration of network evolution (Albert and Barabási, 2002). Similarly, biological network models have been evaluated for their potential to predict the dynamic processes behind the behavior of these systems (Newman, 2003). Among the realms outside of biology, the ‘world wide web’ is a renowned example of an evolving topological model, which uses local connectivity measurements to characterize large scale properties (Albert, Jeong and Barabási, 1999) of a man-made system. In biology, metabolism has been studied extensively to gather information about enzymatic pathways and networks. Metabolism represents a collection of enzymatic reactions and transport routes that convert metabolites into life sustaining molecules and chemical energy. The Molecular Ancestry Network (MANET), and its associated “metabolic MANET” database is a project that traces the evolution of protein architectures in biomolecular networks by linking information from SCOP and the Kyoto Encyclopedia of Genes and Genomes (KEGG). Using phylogenetic reconstructions of protein fold architecture, MANET maps ancestries of enzymes in the pathways of over one hundred metabolic subnetworks, thereby defining evolutionary relationships directly and globally in biological networks. Thus, MANET provides a platform to test hypotheses related to metabolic evolution and study other networks, including those describing biological signaling and molecular interactions (Kim, Mittenthal and Caetano-Anollés, 2006). Metabolic networks create mass and energy, transfer information and specify the fate of a cell. Evolutionarily, they are also highly conserved. According to a survey of 43 organisms from the three domains of life, metabolic networks have been shown to have the same topological scaling properties across life (Jeong *et al.*, 2000). A study of origins and evolution of modern metabolism

identified that the purine metabolic subnetwork was primordial and cradled nucleotide metabolism enzymes with the P-loop hydrolase fold, which is one of the nine most ancient and widely distributed protein fold structures (Caetano-Anollés, Kim and Mittenthal, 2007).

Networks have clarified patterns and processes in structural genomics. The structural topologies of proteins embed evolutionary histories of protein diversification that were uncovered with rooted phylogenies, reconstructed with cladistic principles of sharing of protein folds (Caetano-Anollés and Caetano-Anollés, 2003). A phylogeny of protein architectures described prevalent transformation pathways of domains in architectures and showed that structural SCOP classes of proteins evolved in a specific order, with the alpha/beta class appearing first. An alpha-to-beta tendency was followed by a structural change closely linked to enzymatic function (Caetano-Anollés and Caetano-Anollés, 2003). A hidden Markov model-based phylogenomic analysis of 776 SCOP F structures from 185 completely sequenced genomes confirmed evolutionary patterns of architectural diversification that were conserved and closely related in phylogenomic trees of proteomes at SCOP F and FSF levels (Wang *et al.*, 2006). Another phylogenomic study done in parallel using the same dataset focused on the three domains of life and challenged the accepted eukaryote-archaeobacterial sisterhood by suggesting a closer relationship of Eukarya with Bacteria rather than with Archaea (Wang and Caetano-Anollés, 2006b).

PPI networks that describe the interaction between different proteins have drawn attention to historical patterns of protein combinatorics that directly impact cellular functions. In an effort to model PPIs, a rule-based modeling language named Kappa (<http://www.kappalanguage.org>) was introduced and utilized in tracing histories of biochemical pathways using a notion of trajectory compression, which strictly associates generation of a resultant structure from a pathway to the input (Danos *et al.*, 2012). Kappa has also been applied on kinetic modeling of yeast two-hybrid

data to discover rapid divergence in the space of possible molecular combinations of PPIs, leading to formation of diverse large complexes (Deeds *et al.*, 2012). A related phenomenon termed as “compositional drift” highlights possible drivers behind the phenotypic heterogeneity in cell populations under varying degrees of control. In an analysis of the complex evolutionary pathways followed by ring-like structures, e.g. those of 20S core particle of the proteasome, the problem of deadlocked phase during assembly dynamics was computationally modeled. Weak molecular interactions were found to significantly maximize the robustness and efficiency of assembly for both homo- and heteromeric ring-like structures, owing to intermediate affinities, providing an evolutionary rationale for composition of rings and their efficient synthetic design (Deeds, Bachman and Fontana, 2012).

The analysis of PPIs has also inspired some new classification schemes and databases that take advantage of the mass of information retrieved from structural and functional genomics. Moreover, PPIs can delineate prediction maps based on both structural and genomic information using network visualization and analysis (Skrabanek *et al.*, 2008). A notable endeavor is the establishment of a non-redundant graphical hierarchical classification scheme of 3D protein complexes of the PDB (<http://www.3Dcomplex.org>), which helped understand the evolution of interactions between proteins (Levy *et al.*, 2006). The classification shows that most known structures are multimeric (homo- or heteromeric) and use a limited number of topological arrangements. A different approach on the subject led to the classification of domain interfaces according to their geometry into 6,000 distinct types (<http://scoppi.biotec.tu-dresden.de>). This classification models multidomain protein complexes as multiple binding orientations of interacting domain families, thus allowing evolutionary analysis and database queries of events of genuine gene fusions (Kim *et al.*, 2006). The association of a few protein folds with a multitude

of cellular functions implies a diverse functional repertoire of protein-protein interface architectures (Tuncbag *et al.*, 2008). A database of 8,205 interface clusters showed scarce conservation and usage multiplicity of interface architectures, indicating their towering diversity over folds. A similar database of structures, interfaces and alignments of PPIs (SNAPPI-DB) contains structural data down to the level of atomic co-ordinates, involving multiple data resources such as PDB, SCOP, CATH, Pfam, SWISSPROT, InterPro, Gene Ontology (GO), and protein quaternary structures (PQS) & secondary structure information (<http://www.compbio.dundee.ac.uk/SNAPPI/downloads.jsp>). In another scheme of classification, the FSF-F pair and interaction interface has been used to classify domain-domain interactions (Jefferson *et al.*, 2007). The SCOWLP database portrays a hierarchical classification of 9,334 protein binding regions (PBRs) of 2,561 domain families (<http://www.scowlp.org>), with three additional SCOP sub-levels of F in terms of binding regions, interfaces and contacting domains (Teyra *et al.*, 2008). This allowed to further the study of domain interactome conservation and variability. SCOP is one pioneer database that provides comprehensive taxonomy of structural and evolutionary relationships between protein structures, along with their atomic coordinates, images, interactive views, sequence data and literature references (<http://scop.mrc-lmb.cam.ac.uk/scop/>) (Murzin *et al.*, 1995). A sequence homology-based database search, together with keyword-based searches on Brookhaven protein databank allowed the comparison of homo- and heterodimeric protein-protein interfaces (Guharoy and Chakrabarti, 2007). Homo dimers, for instance, were found to prefer regular secondary structures such as helices and strands in their interfaces, while turns, loops and other substructures were found mostly associated with heterodimers.

The systems biology field of research is rapidly gaining attention with expanding explorations of cell components and their interactions. Biological networks formed by such interactions contain

motifs or recurring interaction patterns, much more often than random networks (<http://www.weizmann.ac.il/mcb/UriAlon/>). For instance, transcription, signal transduction and neuronal networks of Bacteria, mammalian species, and organisms of other superkingdoms seem to make use of network motifs as building blocks. The precise wiring of the components renders biological circuits with special dynamical features such as robustness and enables the circuits to function under varying concentrations of proteins across the cells. Biological systems and networks like these have received primary focus in postgenomic biomedical research and for cataloging and understanding the molecular interactions of a living cell (Barabási and Oltvai, 2004). Numerous cell constituents including DNA, RNA, proteins, and smaller molecules are intricately connected with signaling pathways and regulatory mechanisms, which enables them to adapt to environmental changes. In another approach, genomic and proteomic analyses integrated graph concepts to provide insight into the evolution of cellular network structure and topology (Albert, 2005). An interesting study partitioned metabolic networks into subnetworks and introduced two distinct concepts about substrates (Huss and Holme, 2006). The abundant substrates were termed “currency metabolites” and peripheral substrates were called “commodity metabolites”. Networks of interactions between these two types of substrates contained well defined cores, indicating strong modularity as compared to random networks. However, the modular division was less discernible due to a complicated and eccentric cross-modular connectivity, a feature that apparently holds the key to a robust metabolism. In the post genomic era, systems biology applications now map interactions between biochemical agents onto complex networks (Guimera, Sales-Pardo and Amaral, 2007). Based on the role of agents such as proteins and metabolites in different modules, a network study helped define three enzyme classes: essential, species-specific or both. This provided a classification tool to discover drug targets.

However, metabolic networks are constrained by physical requirements such as conservation of energy and redox, yet they are remarkably robust and evolvable, showcasing a highly variable stoichiometry shaped by the environment (Tanaka and Doyle, 2004; Tanaka *et al.*, 2005). Hence, in order to understand the system of a cell, it is also necessary to extract details of evolutionary links between structure and molecular functions from the network-like hierarchy of environmental interactions (Kim and Caetano-Anollés, 2010). This function-focused phylogenetic analysis used GO terms associated with 38 genomic sequences, revealing a primordial appearance of hydrolases and transferases. Results suggested that biopolymer metabolism preceded macromolecular biosynthesis, challenging the existence of an ancient RNA world.

Networks have been utilized not only in theoretical sciences but also to find solutions to applied sciences problems, especially in genetic medicine. Network medicine is an emerging field that explores molecular relationships among distinct patho-phenotypes that perturb the complex intracellular and intercellular tissue networks (Barabási, Gulbahce and Loscalzo, 2011). Further research in this novel area would help identify disease modules and pathways, new disease genes, drug targets and biomarkers for complex diseases. In another study about biomedical networks, a bipartite network of disorders and disease genes provided a platform to explore the common genetic origin of many diseases (Goh *et al.*, 2007). It was also discovered that a specific set of essential genes encodes hub proteins associated with disorders, as opposed to nonessential disease genes, indicating the existence of distinct disease-specific functional modules. An analogous work on gene networks identified genes responsible for complex human diseases such as autism, inflammatory bowel disease, diabetes, coronary heart diseases and various cancers (Wu *et al.*, 2008). Causative genes of similar diseases were found to share a phenotypic biological module, e.g. a protein complex, a pathway or a subnetwork of protein interactions. Based on these findings,

a method named CIPHER was developed to correlate protein interactions and phenotype networks in order to predict disease genes by using genome-wide information from phenotypically similar diseases. The method identified three putative pairs of cooperative genes in breast cancer. A similar analysis of 506 human protein complexes, which were categorized according to their roles in human diseases, led to a phenome-interactome network using a phenotype similarity score (Lage *et al.*, 2007). A Bayesian predictor then ranked the disease-causing proteins into disorders such as retinitis pigmentosa, epithelial ovarian cancer, inflammatory bowel disease, amyotrophic lateral sclerosis, Alzheimer disease, type-2 diabetes and coronary heart disease. Research in this area also shows that systematic screening of human protein interaction can help understand protein function and associated cellular processes. Using an automated yeast two-hybrid (Y2H) interaction mating method, a highly connected network of 3,186 new interactions among 1,705 proteins was created and investigated to link human disease proteins to regulatory cellular pathways (Stelzl *et al.*, 2005). It also has been discovered that distortion of protein interfaces may lead to development of diseases due to disruption of protein recognition at the molecular level (Shoemaker and Panchenko, 2007). Systematic mappings of PPIs also have revealed global topological features of dynamic interactome networks, which provide insight into disease mechanisms (Rual *et al.*, 2005). An analytical data set called CCSB-HI increased tested binary protein interactions by 70% and revealed 300+ new connections to over 100 disease-associated proteins, an important effort towards advancing the human interactome project.

Proteomics is another vast field where networking tools have been used for illustrative analyses and inferences. A structural classification scheme based on PPIs (SCOPPI) was developed to be applied to the existing SCOP domain definitions in order to devise a comprehensive geometrical classification of domain interfaces based on a distance criterion (Winter *et al.*, 2006). It has been

established that in 98% of proteins a domain pair may occur in only one of the two possible sequential orders (Bashton and Chothia, 2002). These majority cases conserve the domain order, considering that the recombination of the domains occurred only once during evolution, i.e. if two domains combined in both straight and reverse sequential orders, their geometrical relationships would give them different functional properties (Bashton and Chothia, 2002). A study of domain combinations in the proteins of 131 completely sequenced organisms identified domain pairs and triplets that were conserved with a particular functional and spatial relationship (Vogel, Berzuini, *et al.*, 2004). The geometric orientation or spatial separation of the component domains of these domain combinations, which were termed “supra-domains”, was constrained according to their multidomain protein function and respective evolutionary properties. A genome-wide analysis revealed a strong correlation between domain shuffling that causes versatility, and duplication of whole genes that results in abundance (C Vogel, Teichmann and Pereira-Leal, 2005). This correlation supports a nonrandom or stochastic recombination scenario in the evolution of the protein repertoire. A similar view on genetic, protein and metabolic biological webs distinguishes them from static and unweighted networks (Wuchty, 2006). This study, which was defined by the protein family database (PFAM), suggests that protein interactions are probabilistic and are mostly determined by evolutionary conserved single domain interactions. In an extended exploration, less well-defined domains from Pfam-B, and clustered domains from Mkdom after SCOP assignments (MAS), were each compared with the more structured domains from Pfam-A and SCOP (Ekman *et al.*, 2005). The regions of Pfam-B and MAS domains that were non-homologous to Pfam-A or SCOP domains were found to have highly disordered structure, mostly functioning as linkers between structured domains. In a parallel investigation, evolutionary events that distinguish proteins from their ancestors revealed a measure called “domain distance”,

which was defined as the number of domains that differ between two domain architectures (Björklund *et al.*, 2005). This measure highlighted the role of indels over domain repetition and provided a novel method to build evolutionary trees. These analyses also identified how proteins originated from domain architectures.

Protein domains can serve as evolutionary imprints of both protein sequence and structure by consolidating a common ancestry through unification into a smaller set of folding architectures. Rooted phylogenomic analysis of these architectures demonstrated marked heterogeneities in fold distribution (Caetano-Anollés and Caetano-Anollés, 2005). The evolutionary patterns thus observed were related to both protein architecture and organismal diversification. It was revealed that the repertoire of protein architectures and the record of genomic history are hence evolutionarily conserved. A reconstruction of these phylogenomic trees described three epochs in evolution of the protein world, namely architectural diversification, superkingdom specification and organismal diversification (Fig. 1.3). Another notable aspect of this study was the realization of the timeline of architectural diversification in the emerging lineages (Wang *et al.*, 2007). This timeline showcased an initial loss of architectures. It was also postulated that domains not only repeat but also sometimes get recreated, indicating convergent evolution. A domain-tree based investigation also identified the idea of domain architecture reinvention as a common phenomenon (Forslund *et al.*, 2008). This research on Pfam architectures from 96 genomes used maximum parsimony to infer ancestry. Another study focused on proteomes and protein structure uncovered global evolutionary patterns and timelines of discovery of architectures (Caetano-Anollés *et al.*, 2009). In particular, it revealed a biologically complex ancestral proteome that harbored the P-loop-containing triphosphate hydrolase fold. This fold was recognized as the most ancient domain structure. Inevitably, it was highlighted that while sequences have gradually changed over the

course of evolution, contemporary protein architectures can be regarded as molecular fossils.

1.9 Organization of Proteins by Loop and Dipeptide Subunits

Proteins are highly diverse and embed many levels of molecular organization (Fig. 1.2A). At the lowest organization level, *primary structure* specifies the sequence of the monomeric subunits of the polymer and *secondary structure* determines the local folding of the polypeptide chain into regular structured segments. Secondary structure elements include helices (mostly α -helices) and extended chain segments called β -strands. Helical regions locally distort the protein backbone through close-range hydrogen-bonding electrostatic interactions established between main-chain carbonyl (C=O) and amine (N-H) groups typically located 3-5 amino acid residues away in the helical turn. On the other hand, β -strands establish hydrogen-bonding interactions between the main-chain groups distantly located in the polypeptide chain, forming β -sheets in parallel and antiparallel arrangements that often curve into open and closed barrel structures. The ‘helical’ and ‘sheet’ components are separated by rigid stretches of the backbone with no standard secondary structure. These non-regular loop regions (Leszczynski and Rose, 1986) harbor shorter turns and coils that change the direction of the polypeptide chain in space and have functional importance (e.g. β -turns (Dhar and Chakrabarti, 2015)). Notably, α -helices are susceptible to more mutations than β -strands without changing their structure, while diverging more rapidly in sequence than strands of a same structural domain (Abrusán and Marsh, 2016). Both helices and strands are robust to mutational change than regions with no secondary structure, and thus are more designable. This suggests that protein folds resist mutational destabilization by increasing the density of contacts between residues (Tóth-Petróczy and Tawfik, 2014).

The loops are a diverse group of building blocks made of helix, strand and turn segments that are

generally ~25 to 30 amino acid residues long (Söding and Lupas, 2003; Papaleo *et al.*, 2016). Examples include $\alpha\alpha$ -hairpins, $\beta\beta$ -hairpins, and $\beta\alpha\beta$ -elements. Mostly defined as a *supersecondary structures* (Fig. 1.2A), loops collapse into extended flexible or rigid loop-shaped conformations stabilized by van der Waals locks (Berezovsky and Trifonov, 2001; Aharonovsky and Trifonov, 2005). The connector regions of these ‘loop-and-lock’ structures can be exposed to the protein surface and can interact with solvent, ligands, and macromolecules. ‘Loop prototypes (loops) as obtained from ArchDB (Bonet *et al.*, 2014) explain protein function (Feller and Lewitzky, 2012) and are considered determinants of macromolecular dynamics and allostery (Papaleo *et al.*, 2016). While their biophysical properties are driven by yet-to-be discovered evolutionary factors, some recent studies identified combinable (Goncearenco and Berezovsky, 2010, 2012, 2015; Berezovsky, Guarnera and Zheng, 2017) and non-combinable (Alva, Söding and Lupas, 2015) loop motif prototypes that were evolutionarily conserved and thus are likely responsible for the early rise of molecular functions in protein evolution. A first type of loop prototypes combined with others to form active sites that bind cofactors and exert molecular functions (Goncearenco and Berezovsky, 2015). These ‘elementary functional loops’ (EFLs) were obtained through a procedure resembling PSI-BloopST by iterative derivation of sequence profiles from protein coding sequences in complete proteomes scored based on their weights profile positions and according to their information content (Goncearenco and Berezovsky, 2010). Redundancies from overlaps of 10 residues each were then removed by profile clustering using a distance measure that addresses all possible profile-profile alignments. Network visualizations of putative evolutionary connections among EFLs revealed patterns of motif reuse in archaeal proteins (Goncearenco and Berezovsky, 2012). Several network clusters unified significantly known molecular factors such as (i) nucleotide/phosphate cofactor binding domains in ABC

transporters, aminoacyl-tRNA synthetases, and helicases, (ii) metal cofactor binding domains, (iii) methylases and methyltransferases, (iv) transcriptional regulators, and (v) structural repeat building blocks typical of cell surface proteins. This showed that EFLs are ancient structural scaffolds that emerged in protein evolution from shorter protein modules typically 6-9 residues long (Trifonov and Frenkel, 2009; Sobolevsky, Guimaraes and Trifonov, 2013), and were recruited in domain structures to build higher levels of protein organization. Short ubiquitous modules include ‘aleph’ loops and ‘beth’ sequences of ancient ABC cassettes and other regions of ABC transporter proteins. Other highly repeated non-combinable loops present in popular folds represent remnants of an ancient peptide ‘vocabulary’ that combined and accreted to form folded proteins during a primordial RNA-peptide world (Alva, Söding and Lupas, 2015). This comparative approach that identified these supersecondary structures explored local sequence and structural similarities with HHsearch and TM-align, respectively, in protein structural domains unified by <30% sequence similarities. The machine learning tools uncovered biphasic patterns in probability distributions that highlighted high-scoring subdomain-sized fragments. These were clustered as 65 groups, then reduced to 40 groups to avoid redundancies and artifacts from domain boundaries, and finally aligned into folded loop structures that were 9-39 residue long. These loop structures embedded nucleic acid binding fragments with the helix-turn-helix and helix-hairpin-helix motifs and ribosomal protein motifs. Another group of identified loops was enriched in catalytic (P-loop motifs and dinucleotide-binding β - α - β motifs) and binding functions (coordination with metal ions and iron-sulfur clusters). A third generic and widely reused supersecondary structural loop motif has been described with lengths that vary from 35 to 200 amino acids (Nepomnyachiy, Ben-Tal and Kolodny, 2014, 2017). These were used to build networks of protein domains and motifs linked by motif reuse in domains (Nepomnyachiy, Ben-

Tal and Kolodny, 2014). Essentially, these motifs are selected from meaningful all-versus-all SSM or HHsearch alignments. In an improved implementation, candidate motifs were identified using an efficient dynamic programming algorithm that uses a two-step function to find widely reused contiguous non-overlapping segments of widely different lengths, called ‘themes’ (Nepomnyachiy, Ben-Tal and Kolodny, 2017). Interestingly, the increase in reuse with decreasing theme length also followed a power law, which indicates a significant biased distribution of these motifs in proteins. All these approaches indicated that (i) supersecondary motifs are identified by sequence and sometimes structural similarities, and may or may not carry evolutionary relationship, (ii) motif recurrence across proteins is driven by biological function, and (iii) motifs form complex patterns in proteins are suggestive of an interplay of divergent vs. convergent evolution, perhaps subject to the existence of rearrangements, duplications, and divergences. Loops and themes, can thus be considered ancient evolutionary building blocks.

A study focused on the accretion of proofreading and specificity domains of aaRSs and their interaction with evolving tRNA (Caetano-Anollés, Wang and Caetano-Anollés, 2013). It was found that the most ancient aaRS domains harbor archaic dipeptidase and ligase activities, indicating that the primordial peptides and polypeptides that were originally produced as a result conditioned the amino acid makeup of modern proteins. This hypothesis was tested by observing the sequential overlap of dimers in the amino acid sequence of proteins. These ‘dipeptide’ constituents were considered as relics of ancient amino acid composition. Each dipeptide conforms to one of ~400 pairs of two sequential residues or ‘peptide bond types’ that make up a protein sequence. The results showed that the dipeptides of protein domains appearing in evolution before the first anticodon-binding domain were significantly enriched in aaRS editing amino acids. Also, results drawn directly from phylogenomic, and structural information uncovered a hidden link

between the emergence and expansion of the genetic code and protein flexibility. This linkage provided understanding to one of the most challenging problems prevailing in biology. The study used protein domain definitions, RNA substructures and dipeptide signatures as ‘basic’ evidential statements within the Popperian framework and regarded these as agreed upon conjectures of perceived similarities, strengthened by reciprocal illumination, and defining useful homologies, trees, and timelines. Since dipeptide synthases require charged tRNA to function, it was conjectured that archaic aaRSs capable of aminoacylation and peptide bond formation cooperated to perform primordial protein biosynthesis before the existence of the ribosome (Caetano-Anollés, Kim and Caetano-Anollés, 2012). To support this conjecture, the expected impact of dipeptide synthesis on the peptide bond (dipeptide) makeup of protein structures was investigated. Evolutionary heat maps describing the amino acid and dipeptide compositions of 1,475 domain domain structures of different age confirmed the existence of non-random patterns of amino acid and dipeptide compositions along the phylogenetic timeline of domains and revealed unique signatures of amino acid and dipeptide use in domains. Dipeptides thus provide unique signatures of fold structure (Nakashima, Nishikawa and Ooi, 1986) with predictive power matching that of domains (Roy *et al.*, 2009).

1.10 Scale-free and Sister Explanations of the Growth Patterns of Biological Networks

The functions of genes are expressed through the resulting proteins or functional RNA molecules. In evolution, genes produce new genes by duplication, mutation, recombination, fusion and fission. These diversified genes express new proteins, which sometimes display new biological functions. Research has attempted to quantitatively describe the origins of these processes, which increase molecular complexity over the course of evolution through pathways of protein domain

organization (Chothia *et al.*, 2003). Initial network studies explored these processes with protein sequences from two separate sets containing 7 and 40 genomes, respectively, in order to study their evolutionary units, the protein domains (Apic, Gough and Teichmann, 2001b, 2001a). A total of 624 and 783 SCOP FSFs present in each genome set made a network of 585 and 1,307 pairwise combinations, respectively. Interestingly, only a few FSFs occurred in most of the combinations, following the pattern of a scale-free network. A similar theoretical model of protein domain networks was constructed that connected proteins when they shared a domain (Nacher, Hayashida and Akutsu, 2006). The degree distribution of the resulting network showed superposition of two power-law functions, with exponents of -1 and +1, respectively. This pattern, which was coined as “scale-free mixing”, was attributed to mutation and duplication of domains. Another study gathered domain data of different organisms from ProDom, Pfam, and Prosite domain databases (Wuchty, 2001). It was revealed that the resulting protein domain networks exhibited scale-free or small-world topologies, i.e., a high degree of local clustering and few long-distance links. Moreover, the networks showed that the extensive connectivity among domains reflected the evolutionary complexity of the various organisms.

Preferential attachment is a principle popularly used to explain patterns of growth in evolving networks. Mathematically, a network is a graph with vertices (also termed nodes or actors) and edges (links or connections) with functions or parameters mapped to both. Networks with topology as complex as genetic networks or the World Wide Web have been attributed with a power-law based scale-free degree distribution. In other words, expansion of such large networks follows a self-organizing phenomenon that adds new vertices in a preferential manner to already well-connected hubs, despite of the properties of individual vertices or systems (Barabási and Albert, 1999). The connectivity of vertices in such networks decays as power law, i.e. the probability of

having at least k connected vertices for each vertex decays as $P(k) \sim k^{-\gamma}$, a feature distinct from those of random networks. In contrast, the random graph model of Erdős and Rényi provides a view that is tangential to the scale-free model (Erdős and Rényi, 1960; Bollobas, 1985). The probabilities of connectivity of random graphs follow a Poisson distribution with $\lambda > 1$. A special case of the random model is the small world model (Milgram, 1967; Kochen, 1989) introduced by Watts and Strogatz (Watts and Strogatz, 1998), also referred to as ‘six degrees of separation’ (John, 1990). In this model, vertices connect to their nearest and a next-nearest neighbor, and then reconnect to a long-range vertex chosen at random, eventually decreasing the path distance between the vertices. In general, both the small world and random models diminish the probability of a vertex having higher connectivity. In other words, under these models, $P(k)$ does not decrease exponentially with k , and thus highly connected vertices practically vanish. This aspect derives from the assumption of uniform probability of connectivity among vertices, contrasting to the patterns of preferential connectivity. While scale-free networks come under the umbrella of small world networks, sometimes regarded as ultra-small (Cohen and Havlin, 2003), these networks showcase a power-law tail, with a close-to-zero chance of occurrence of highly connected vertices (Barabási and Albert, 1999).

Scale-free explanations of networks are widely prevalent in both the technological and natural realms. A prominent property of large scale-free networks, such as evolving biological networks, is resilience against random breakdowns (Dorogovtsev and Mendes, 2002). While this error tolerance (ET) is primarily a biological feature that allows even simple organisms to persist drastic pharmaceutical or environmental interventions because of a robust metabolic network, ET has also been associated to artificial networks that are inhomogeneously wired or are scale-free. Nevertheless, ET is accompanied with vulnerability to targeted attacks on vital hub nodes, for

instance, in global communication networks (Albert, Jeong and Barabási, 2000). This phenomenon has also been supported by simulated annealing optimization on heterogeneous nets, which were shown to cluster around a well-defined narrow domain (Solé and Valverde, 2004). Such optimization of the entropy-noise space limited the scope of possible complex networks to those networks that for example harbor evolution. A review of cellular networks, including metabolic and protein networks as well as expression profiles, revealed decaying power-law patterns governed by generic laws and organizing principles (Nacher and Akutsu, 2007). The internet, which has an apparently random topology, also seems to exhibit power law-like expansion (Faloutsos, Faloutsos and Faloutsos, 1999). An extensive review of coupled dynamical systems described the neural network of the worm *Caenorhabditis elegans*, indicating features of a small-world or self-organizing network model (Huang, Jedynak and Bader, 2007). This study also investigated yeast two-hybrid screens of pairwise physical interactions between proteins revealing that the membrane proteins had higher false-discovery rates than signal transduction proteins. Interestingly, these originally highly clustered but regular networks rewired to introduce disorder with characteristic small path lengths, mimicking the ‘six degrees of separation’ phenomenon. Statistically, the degree distribution patterns favored a scale-free model of yeast and a truncated power law models of both the worm and the fly, as opposed to an Erdős–Rényi random network model. Evidently, small-world coupling enhances signal-propagation speed and synchronization, for instance, in spread of infectious diseases (Watts and Strogatz, 1998). Further clustering tendency of small-world networks leads to a modular property of the network where cohesive groups of nodes interact with each other as if they are nodes themselves, i.e., strong intra-links within a module with weaker inter-links between modules. Modularity usually mirrors scale-free property of a network in a symmetric manner, i.e., in most cases a

network's connectivity distributions can either have power law features or show modularity. However, although rarely but certain networks, especially in biology, display qualities where both preferential attachment and modular behavior can be derived from network progression, a phenomenon becoming popular in biological networks and referred to as hierarchical modularity (Ravasz *et al.*, 2002; Caetano-Anollés *et al.*, 2019). These various forms of network structure provide a morphospace in which instances of a growing network can be mapped via indices of major network features such as scale-freeness or heterogeneity, randomness or robustness, and modularity or clustering (Fig. 1.9), which allows visualization of multiple network features in a single 3D plot (Caetano-Anollés *et al.*, 2019).

Research has also uncovered networks of backbone importance, especially in biology, whose growth might not be fully explained by the preferential attachment principle. In large non-biological networks (e.g. the internet), modular degree distributions have been demonstrated to be exponential rather than power-law (Li *et al.*, 2004, 2005; Tanaka, 2005). Another study used Shannon entropy measures to categorize evolving networks (specifically social networks) as assortative and correlated or anti-correlated and scale-free (Johnson *et al.*, 2010). Similar results were obtained in a study of evolutionary optimization of dynamic processes on so-called natural scale-free interaction networks, where low scaling exponents and increased randomness were attributed to positive assortative classification (Payne and Eppstein, 2009). An extensive study of metabolic networks, which account for a major section of biological networks, was conducted to examine presence of power-law behavior (or lack thereof) (Tanaka and Doyle, 2004; Tanaka *et al.*, 2005). The insight into the highly optimized and error tolerant biochemical interactions of the metabolic networks explained why scale-free properties fail to correctly describe metabolism. In general, scale-free networks fragment into unlinked subnetworks on removal of the few high-

degree nodes or hubs. However, metabolic networks were shown to deviate from such behavior, where the biosynthetic pathway graphs remained intact even after removal of all high-degree carrier nodes. Remarkably, the stoichiometry of metabolic networks showed highly organized tradeoffs and tolerances when influenced by the environment. This indicates that the metabolic networks harbor a more robust architecture based on a highly structured modularity aspect that is scale-rich rather than scale-free. Thus, metabolic networks were not regarded as “self-organized” but as “self-dissimilar” (Li *et al.*, 2005), with exponential metabolite degree distributions at the module levels. Another critical review of preferential-attachment in biological networks focused on the relationship between power-law and log-normal distributions (Arita, 2005). It was observed that the natural selection process of evolution is not characterized by power-law distribution. A similar review of PPI networks qualified the apparent scale-free properties (Tanaka, Yi and Doyle, 2005). Despite the presence of prominent high-degree hubs, in certain cases the node-degree distribution of PPIs did not follow a power-law. Moreover, sampling and noise ambiguities were identified in the previously used methods. An interesting investigation on human cells about frequency of proteins composed of a certain kind of domain fitted an evolutionary model as an alternative to power-law (Nacher *et al.*, 2009). The investigation found that the protein-count distribution followed a decreasing exponential function instead of preferential attachment. The alternate evolutionary model was based on growth and copy mechanisms. In a related work involving the field of metabolomics, the networks of metabolites of *Escherichia coli* were subjected to various induced stresses to demonstrate erratic modifications in the patterns of modularity and rewiring (Aziz *et al.*, 2012). The behavior of *E. coli* metabolome under the stress factors supported biphasic or hourglass-like patterns, which were divergent from any scale-free behavior.

These observations have led to realization of a strange phenomenon in biology, growth of modules within modules in a power law fashion, which at the same time is relatable to a prophetic historical view that can be termed as a blueprint of growth patterns of timed network systems. Interestingly, the 2000-year-old Strasbourg papyrus attributed to Empedocles of Akragas (ca. 495-435 BC) appears to summarize the double tale of module generation and its dynamic nature (Fig. 1.10). The papyrus fragment ensemble fr. **a**, for instance, presents crucial principles that are common to modern evolutionary biology, which become gradually and sometimes cryptically evident (Caetano-Anollés *et al.*, 2019). Lines 258-260 specify how growing things establish a *hierarchy* of wholes unified from integrated parts to make up what can only be interpreted as lineages, “*For all these things are equal and alike in age but each rules separate domains*”. Lines 261-266 extend the concept of lineages of a hierarchy (a tree) to lineages of an evolving network, “*But these are what there is, and running through each other they suffer change continual but always are alike*”. Lines 270-274 indicated that those lineages of the network then “blossom” into species, “*as trees, as men, as women too, as beast, as birds, as fish that waters rear*”. Lines 285-287 later restates how parts unify into wholes, “*through one another race, and, roaming, visit other places constantly*”, making explicit time trajectories of evolutionary recruitment that make network structures. Finally, lines 293-298 describe how unification “*augments with larger form life’s union and increase*” to generate a wealth of organismal diversity, which is then catalogued. Remarkably, these interpretations of recurrent arguments gradually advance two concepts linked to the double tale, accretion, and diversification, which are emergent features in evolutionary and systems biology. A network paradigm can be used to illustrate how the biphasic model of module generation, the ‘double tale’, explains the evolutionary emergence of both modularity and preferential attachment in evolving networks (Fig. 1.11). Modular networks harbor ‘communities’,

groups of nodes that are more densely connected with themselves than with the rest of the network. Modularity usually counterbalances power-law behavior of networks, in which case networks generally grow by preferential attachment of new nodes to highly connected nodes (Jeong *et al.*, 2000), but both properties can coexist in networks when modules coalesce hierarchically (Ravasz *et al.*, 2002). A hierarchy is a “*system that is composed of interrelated subsystems, each of the latter being, in turn, hierarchic in nature until we reach some lowest level of elementary subsystem*” (Simon, 1991). Hierarchy in the context of networks is the fractal-like reuse or embedding of simpler network modules into modules of higher complexity. Thus, lower-level network modules are subordinated by an “*authority relation*” to higher network modules (Simon, 1991). Hierarchical modularity can impart numerous benefits to a system over monolithic integrated design. For example, reusability: the design of a module occurs once but it can be reused many times, while every monolithic system must be built from scratch. Simon’s famous parable of the ‘Hora’ and ‘Tempus’ watchmakers describes the cost benefits of building, modifying or updating modules (Simon, 1991). In network biology, hierarchical modularity can increase evolvability: the ability to adapt and innovate (Kashtan and Alon, 2005; Kashtan, Noor and Alon, 2007; Lorenz, Jeng and Deem, 2011), and robustness: the maintenance of function in spite of internal or external perturbation (Kitano, 2004), while decreasing the costs of establishing network connections (Mengistu *et al.*, 2016). Finally, hierarchical modularity can also improve the speed, stability and quality of information transfer through networks (Solé and Valverde, 2004). In evolving networks, nodes organize into parts of a growing system and their interactions become established links between them (Fig. 1.11). The ‘double tale’ takes advantage of the stability that accretion of nodes and links provide to the system. The larger the number of links, the more cohesive and stable is the structure of the evolving network. In a first phase of the ‘double tale’ or

a biphasic model, the accretion of nodes and links is non-specific, but with time it gives rise to highly connected groups. These groups of nodes become modules when links are stabilized by competitive optimization. In a second phase, the emergent modules of the network diversify and coalesce by combination into higher level modules of the network structure. This cyclic process gives rise to a fractal-like pattern of connectivity, hierarchical modularity, which is gaining popularity to explain growing biological systems (Ravasz *et al.*, 2002; Caetano-Anollés *et al.*, 2019). To examine the factors and effects of this theory of the emergence of hierarchical modularity, we traced the history of biological networks of protein domain organization, elementary functionome (EF) and dipeptides. We focused on these networks as they unfold at comparative, biologically entrenched timescales (billions of years), and thus provide a layered ‘parallel universe’ like vertically-arranged-canopies scheme of adaptability and transference of molecular structure and function.

1.11 The Focus of this Study

Here we map the coevolutionary history of the domains and domain combinations of proteins, defined at the fold superfamily (FSF) level of SCOP classification (Murzin *et al.*, 1995), loop prototypes (loops) as defined in ArchDB (Bonet *et al.*, 2014), and dipeptides. Our study investigates the evolutionary mechanisms and the rationale driving the history of networks from the perspectives of protein domain organization, loops, dipeptides. It also analyzes the dynamics leading to an adaptation of the structure of these networks to scale-free properties. We also explain the underlying processes responsible for such network structure. To conduct this investigation, we highlight the cooption of protein domains as they interact with other domains throughout a timeline of protein evolution. For this purpose, a subset of 6,162 protein domains and multidomains was

taken into consideration. Our network model consists of five distinct composition-based and topology-based ‘operational’ criteria for defining nodes and links of the networks: 1) domain and multidomain (domain organization) composition, 2) domain and multidomain pairs decomposition, 3) domain and multidomain adjacent-pairs decomposition, 4) splicing of domain pairs from domain organizations, and 5) splicing of adjacent domain pairs from domain organizations, which we carefully define. The evolutionary mappings (based on the ages) of the protein domain organization and growing interactions revealed the history of an evolving biological network. This evolutionary timeline stemmed from a phylogenomic analysis (Wang and Caetano-Anollés, 2009) and was demarcated by 169 unique time events. In case of loop prototype networks, 1,937 loops and 1,442 fold families (domains) were mapped with 228 unique time events. Also, abundance data of course 400 dipeptides for domains and genomes was considered over 225 unique time vents. These time events (or time slivers) were mapped onto the evolving networks to depict evolving network configurations from the origin of proteins to the modern protein world. The history of these networks was visualized using Pajek (Mrvar and Batagelj, 2016) to reveal distinct patterns of connectivity among domains and multidomains, loops and domains, and dipeptides, and highlight variations of inherent characteristics of domains along the timeline. Particularly, we discover an interesting pattern of highly concentrated connectivity associated to the domains and loops of either the earliest or the post-middle era of domain evolution, while sparse connectivity exhibited by the domains of the middle period of the evolutionary timeline. This observation provides evidence for a biphasic (Mittenthal, Caetano-Anollés and Caetano-Anollés, 2012) or hourglass (Caetano-Anollés *et al.*, 2011) pattern of connectivity in protein domain, loop and dipeptide networks. This pattern seems to favor the ancient and pre-modern entities to form new combinations (by fusion) or to fragment into new

structures (by fission) later in evolution. In turn, the domains, and loops of the middle period of the timeline struggle for dominance. We also dissect two prominent aspects of the evolutionary process of recruitment, the ability of domains, loops, and dipeptides to act as 'donors' and 'acceptors'. These networks were also evaluated using scale-free, random, and modular network models to interpret the gradual evolutionary selection of one model in favor of the other. We thus reveal remarkable pattern of emergence of a hierarchical modularity of structural and functional cooption and pairwise recruitment through exploration of network connectivity in the domain-multidomain, loop-domain, and dipeptide-dipeptide mappings that developed over the course of evolution.

1.12 Introductory Figures

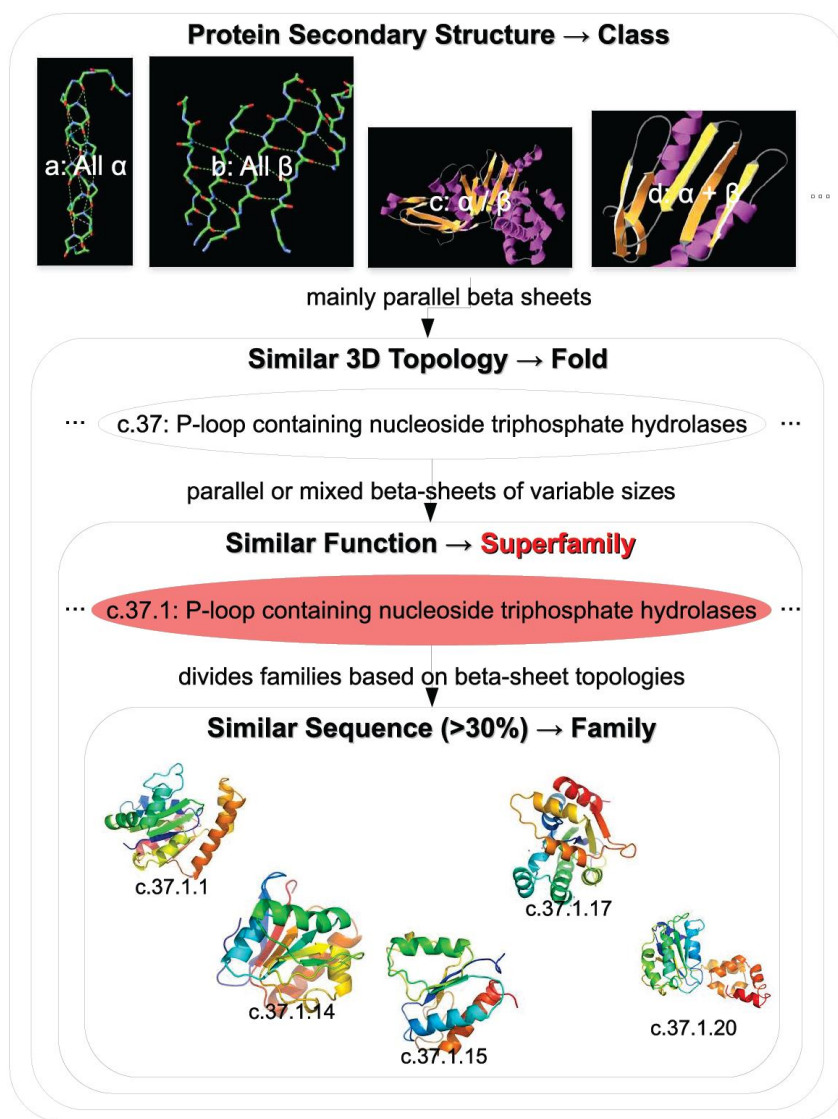


Figure 1.1: SCOP classification scheme. The data set used in this study comprised of domains (and their combinations) defined at the superfamily level. Image sources: *Protein Secondary Structure Types* [3D Model]. (2012). Retrieved May 17, 2018, from: <http://www.proteinstructures.com/Structure/Structure/secondary-structure.html>; *Sample c.37.1 families* [3D Model]. (2009). Retrieved May 17, 2018, from: http://supfam.mbu.iisc.ernet.in/scop_c.html.

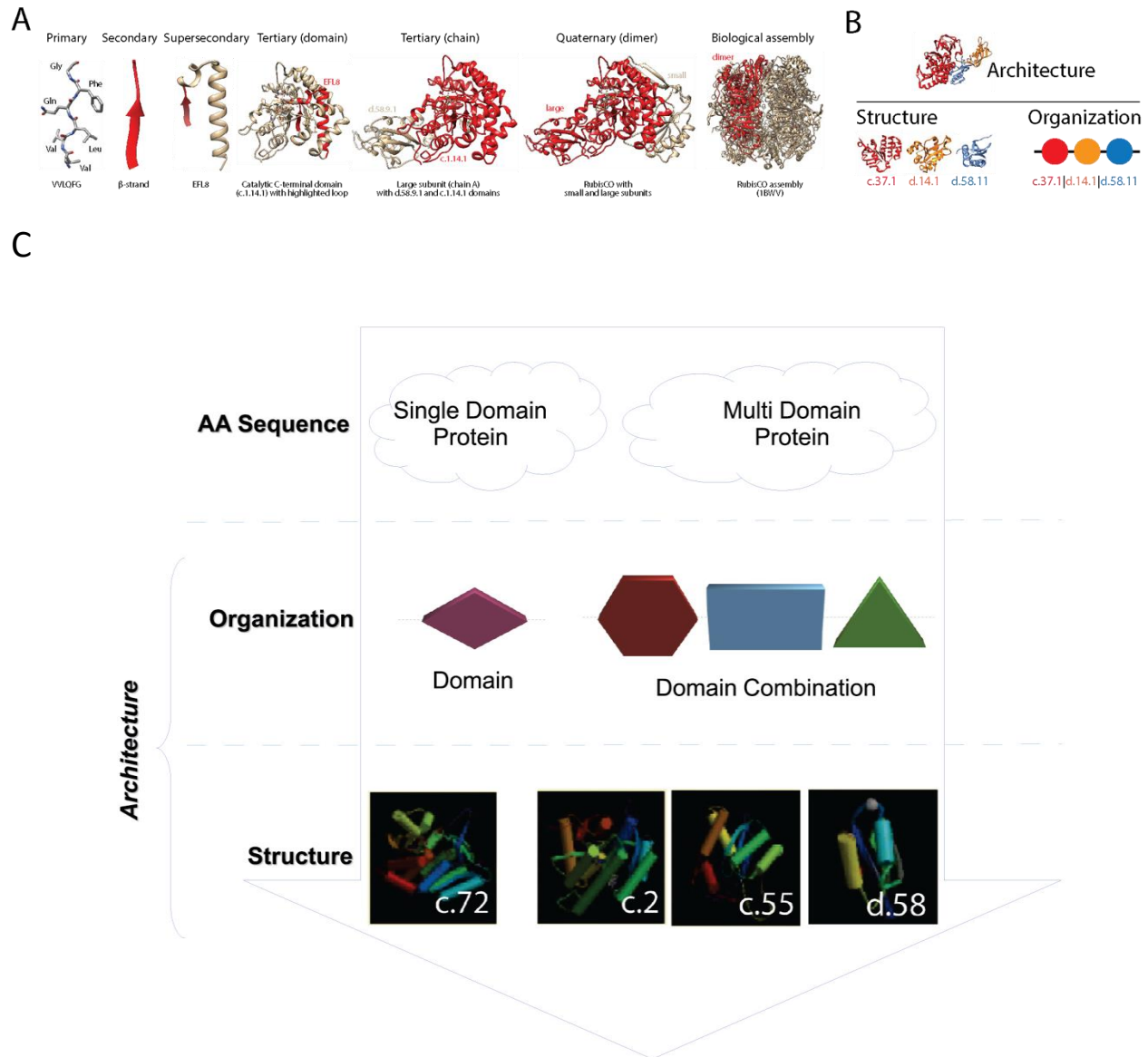


Figure 1.2: Protein makeup and domain architecture. **A)** Levels of molecular organization embedded in a protein biological assembly, the ribulose-1,5-bisphosphate carboxylase enzyme (RubisCO; PDB entry 1WDD). **B)** Elongation factor G from *Thermus thermophilus* (1DAR) illustrates how a domain architecture describes both the structure and organization of its structural domain components. **C)** Another view of how a protein can be divided into component parts called domains, which are structural, functional and evolutionary units of proteins. Each domain has its own folding structure, which defines how sequence of a domain in the protein folds in 3D space at F, FSF, or other levels of architectural hierarchy. In multidomain proteins, several domains ‘organize’ along the amino acid sequence of the protein, delimiting how domains combine with other domains in the polypeptide chain. The arrangement of these domain structures is called ‘architecture’. Therefore, domain architecture defines both the organization and structure of proteins. Image source: (Wang and Caetano-Anollés, 2009).

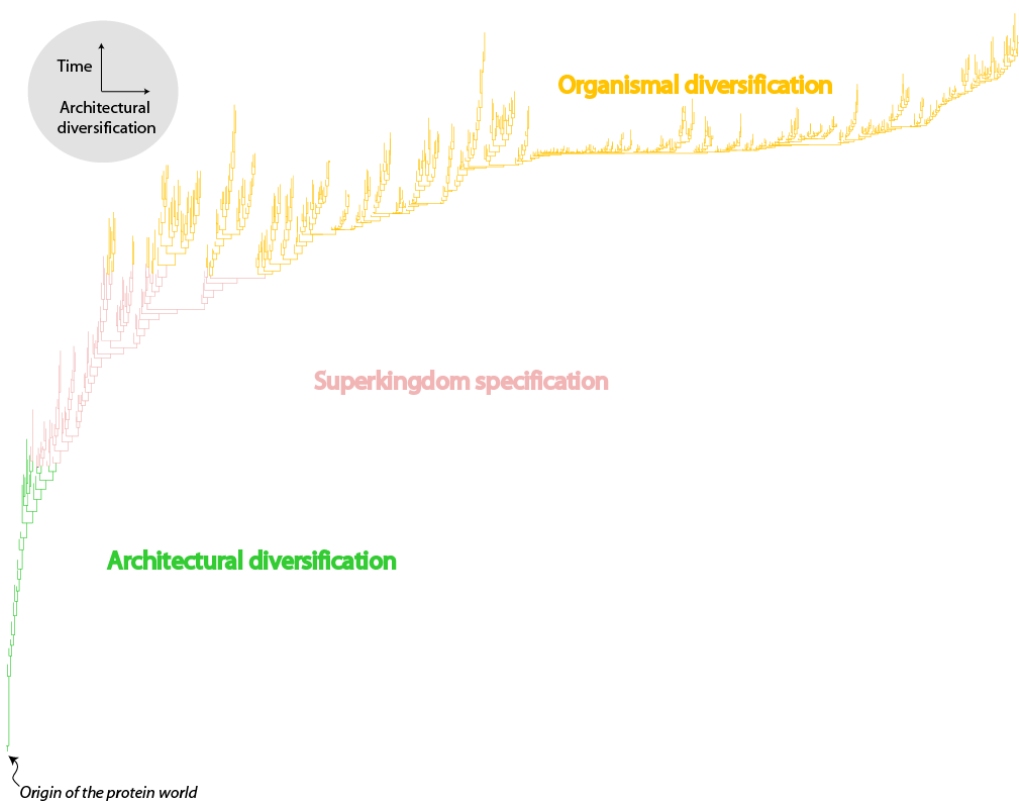


Figure 1.3: Phylogenomic tree of domain organization. The tree of domains and domain combinations was generated by phylogenomic analysis using maximum parsimony. The three colors correspond to the evolutionary epochs of the protein world. Image and coloring scheme source: (Wang *et al.*, 2007; Wang and Caetano-Anollés, 2009).

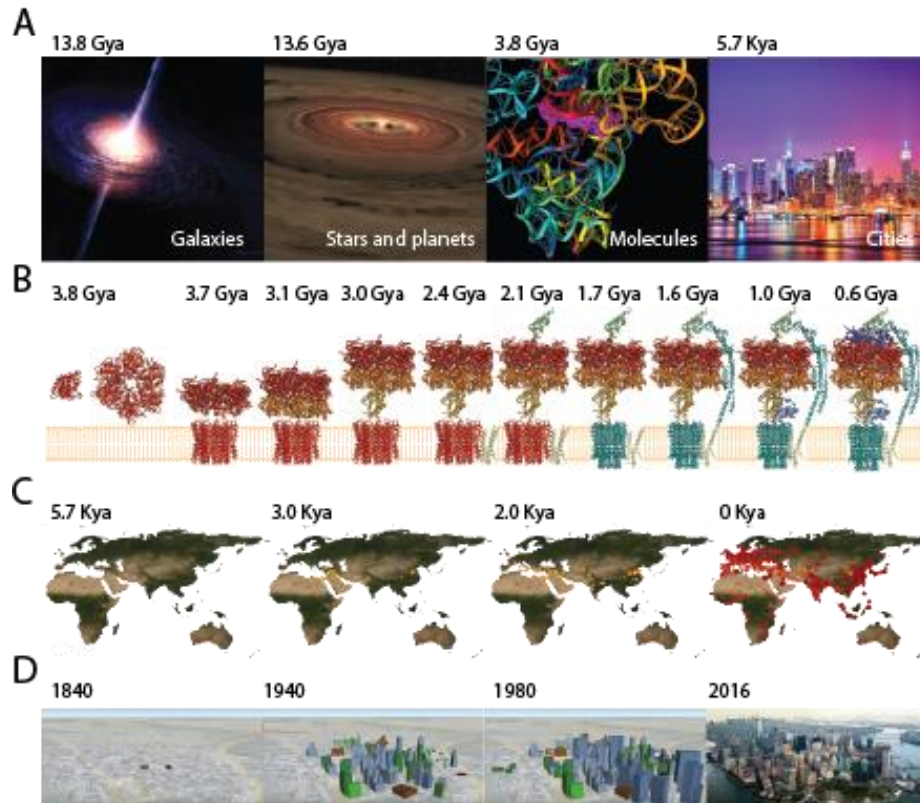


Figure 1.4. Accretion appears universal. (A) Phylogenetic and polyphasic analyses reveal the evolutionary growth of galaxies, stars, planets, and macromolecules. Archeology and historical records uncover the growth of cities. Time of origin is given in billions or thousands of years ago (Gya and Kya, respectively). (B) Evolution of the molecular structure of the F₁/F₀ ATP synthase complex that is involved in bioenergetics of the cell, with structural domains of proteins colored according to their evolutionary age, from red (early) to blue (late). (C) Accretion of cities in Afro-Eurasia and Oceania during 5,700 years of history. (D) Growth of high-rise buildings of Lower Manhattan (floor-level visualizations from CubeLease, <https://www.youtube.com/watch?v=Vc5r8osWauU>).

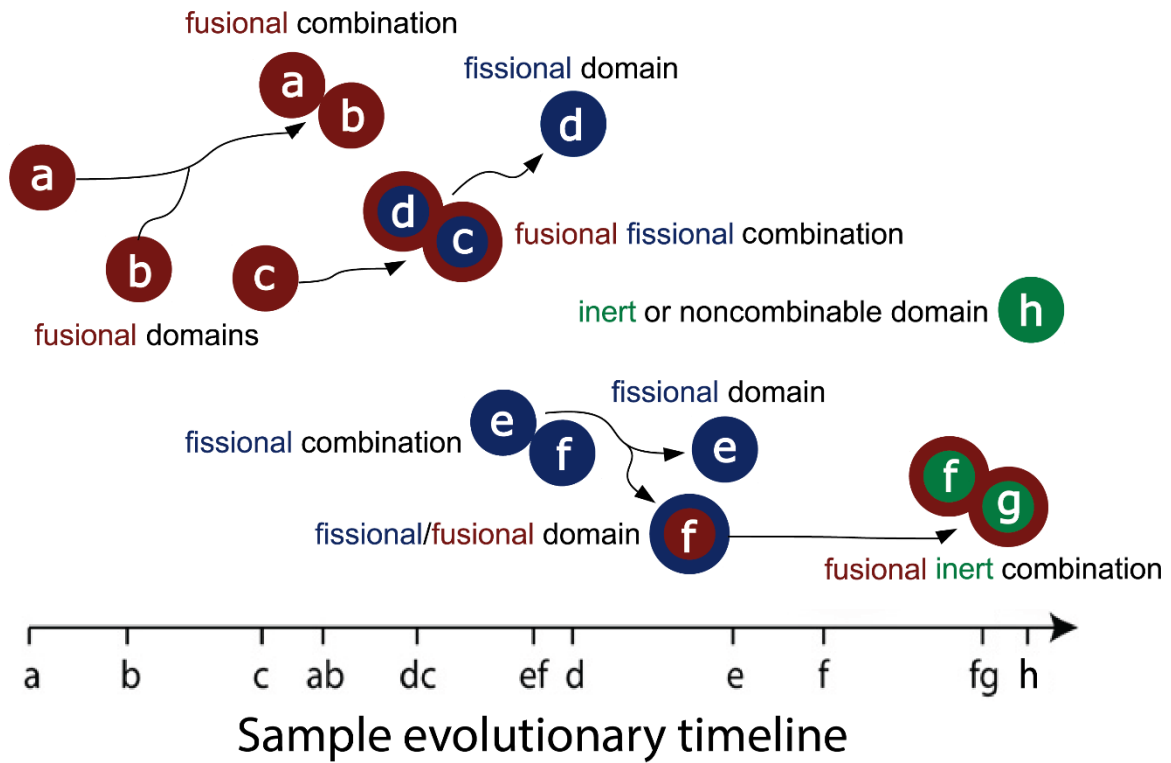


Figure 1.5: Fusion and fission processes. Fusional domains combine to form fusional combinations and fusional/fissional combinations. The latter may eventually yield fissional domains. A fissional combination can split into fissional domains and fissional/fusional domains. The latter may merge into an inert domain combination later. The inert combinations and noncombinable domains do not proceed in the combinatorial play. The various attributes are color-coded as: Fusion – red, fission – blue, and inert – green.

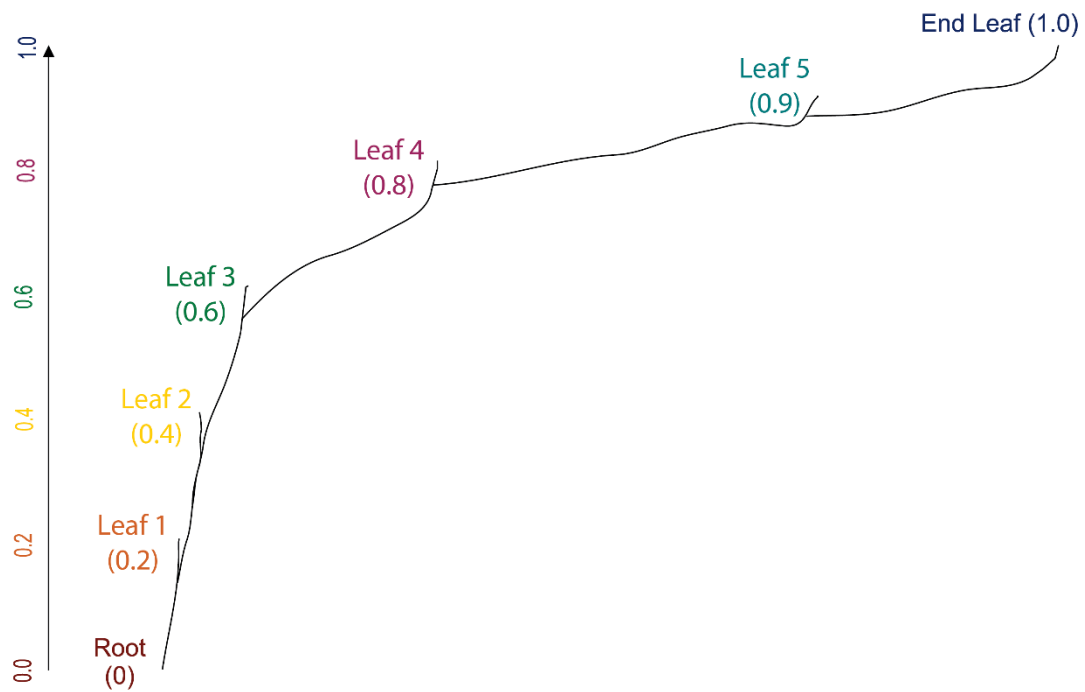


Figure 1.6: Phylogenomic trees have information to produce an associated timeline. Trees of domain organization are highly unbalanced. This property of trees allows to assign relative ages to their terminal leaves. The values in parentheses for each leaf label are evolutionary ages, i.e., normalized node distance (*nd*) values. Note that the 'root' and 'end' leaves have values 0.0 and 1.0, respectively. The timeline at the left is directly inferred from the tree. Labels along the timeline and those in the tree are color coded as a gradient to depict progression of time.

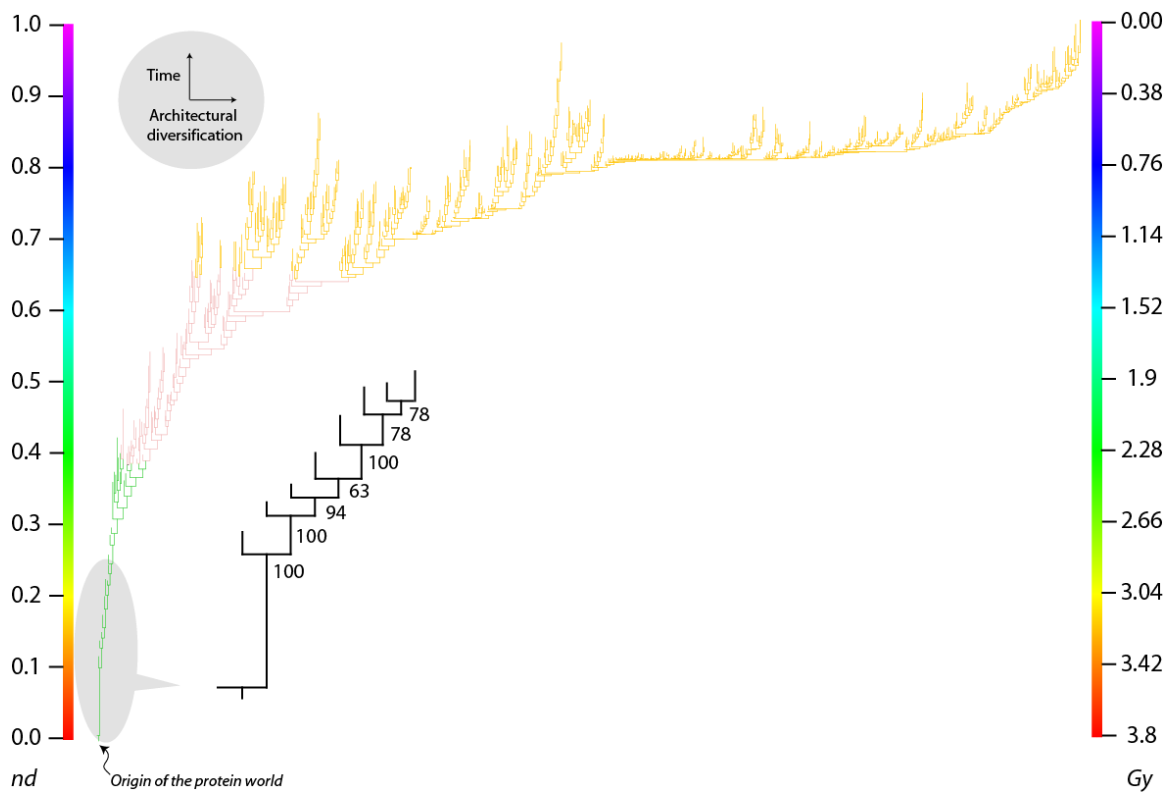


Figure 1.7: Phylogenomic tree of domain organization and associated evolutionary timelines. The rooted phylogenomic tree enabled extraction of timelines of evolution of protein domain organization (Wang and Caetano-Anollés, 2009). The inset shows the values for the first few domains from the phylogenetic distance matrix. The quotient of the phylogenetic distances to the diameter of the tree (the maximum distance) gave the normalized node distance values, nd , displayed on a scale from 0.0 to 1.0 on the left. The heatmap color gradient of the nd timeline corresponds to the age of domain organization. A reference timeline derived from a molecular clock of domain structures calibrated the relative timeline of node distances according to a geological time scale that began billions of years (Gy) ago (Wang *et al.*, 2011). This timeline is shown on the right. The RGB gradient colors were mapped to the corresponding time markers of both nd and Gy timelines. The colored branches of the tree represent the three evolutionary epochs of the protein world (Wang *et al.*, 2007).

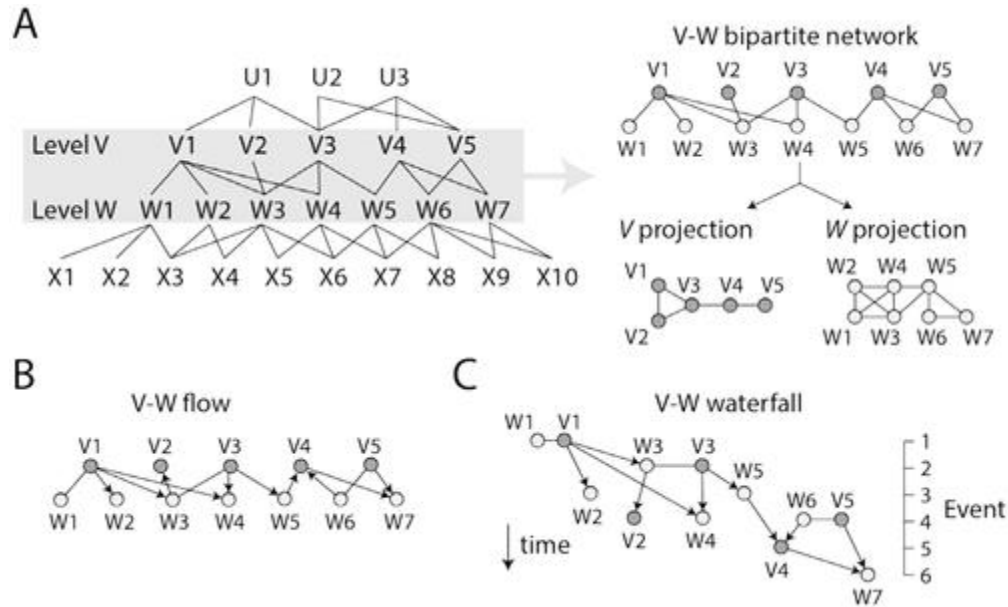


Figure 1.8. A bipartite network view of levels of organization. **(A)** Any system of interacting entities describable with networks can be dissected into a hierarchical system with nested entities defining different levels of organization (eg, U, V, W, and X). Network interactions that are tightly knit generate modules, which enable the functional activities of the system. A bipartite network makes explicit the relationship between any 2 levels of organization when it is dissected into its 2 one-mode projections. One projection describes how higher-level entities link lower-level entities to each other. The other describes how lower-level entities link higher level entities to each other. As an example, a bipartite network describing interactions between entities of the V and W levels is shown in the right together with its corresponding V and W projections. For simplicity, links are left unweighted. **(B)** The V-W bipartite network is transformed into a flow hierarchy when some or all connections are described as arcs pointing in the direction of time. **(C)** The flow hierarchy becomes a waterfall diagram when the ages of nodes are treated as “time events” and are used to reorganize the network in the direction of time.

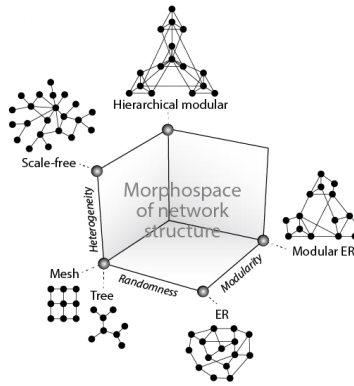


Figure 1.9. Morphospace of network structure. Toy examples of typical graphs describe archetypes of the phenotypic landscapes. In this morphospace, Erdős-Rényi (ER) random graphs transform into regular graphs by decreasing randomness, or into modular ER graphs by increasing modularity. Hierarchical modular structure requires both increasing modularity and heterogeneity and decreasing randomness.



Figure 1.10. Indexed translation of ensemble fr. a of *P. Strasb. Gr.* Inv. 1665-6, with its two columns. The scan shows the reconstruction of the ancient papyrus from fragments rendering Empedocles' *On Nature*, lines 262-300. The scribal stoichiometric sign Γ is of significance as it indicates that the line corresponds to v. 300 of *Physika* Book I of Simplicius. Text lines of theses 2 and 3 of the translation (Janko 2004) are indexed with numbers and highlighted in black when they are part of fragment (fr.) ensemble a. Please refer to Janko (2004) for the original Ionic text of the translation. The inset shows a picture of a stater coin of Aegina ca. 480 BC (194 grs.; British Museum; Head 1888) with a sea turtle on one phase, the creature sacred to Aphrodite, and a three-legged triskelion on its reverse. Public domain images obtained from wikicommons and google.

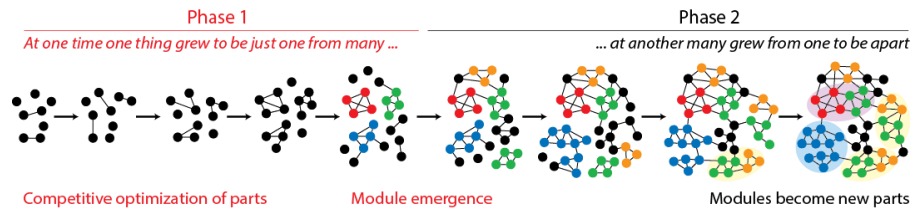


Figure 1.11. A generic biphasic model of module creation illustrates the emergence of network structure in evolution. Nodes and links of the network are parts of a growing system of entities and interactions. The larger number of links, the more cohesive and stable is the structure of a subnetwork. The rise of modularity during phase 1 results in small highly connected subnetworks. These subnetworks become modules, which in phase 2 coalesce by combination into hierarchical modules of network structure (highlighted with shades of yellow and blue). The model is inspired by the work of Mittenthal et al. (2012).

1.13 Introductory Tables

Table 1.1. Comparison of SCOP, SCOPe, SCOP2, CATH and DALI classification systems of structural domains. Level 1 describes the gross primary secondary structure content of structural domains. Level 2 describes the orientation, connectivity, and topology of secondary structures. Level 3 describes structures sharing common ancestors with similar biological function. Level 4 describes structures with sequence similarities of >30 for SCOP, ≥ 35 for CATH and <25% for DALI (using non-redundant PDB25).

Level	SCOP 1.75 (June 2009) (hierarchical)	SCOPe 2.07 (January 2021) (hierarchical)	SCOP2 1.0.5 (April 2021) (hierarchical)	CATH 4.3 (July 2019) (hierarchical)	DALI (April 2021) (threshold-delimited)
1	Class (7)	Class (7)	Class (7)	Class (6)	–
2	Fold (1,195)	Fold (1,232)	Fold (1,500)	Architecture (41) Topology (1,390)	–
3	Superfamily (1,962)	Superfamily (2,026)	Superfamily (2,681)	Homologous Superfamily (6,631)	–
4	Family (3,902)	Family (4,919)	Family (5,627)	Family (32,388)	Chains (21,214)

1.14 Introductory References

Abrusán, G. and Marsh, J. A. (2016) ‘Alpha helices are more robust to mutations than beta strands’, *PLoS computational biology*. Public Library of Science San Francisco, CA USA, 12(12), p. e1005242.

Aharonovsky, E. and Trifonov, E. N. (2005) ‘Protein sequence modules’, *Journal of Biomolecular Structure and Dynamics*. Taylor & Francis, 23(3), pp. 237–242.

Albert, R. (2005) ‘Scale-free networks in cell biology’, *Journal of cell science*. Company of Biologists, 118(21), p. 4947.

Albert, R. and Barabási, A. L. (2002) ‘Statistical mechanics of complex networks’, *Reviews of modern physics*. APS, 74(1), p. 47.

Albert, R., Jeong, H. and Barabási, A. L. (1999) ‘The diameter of the world wide web’, *Arxiv preprint cond-mat/9907038*.

Albert, R., Jeong, H. and Barabási, A. L. (2000) ‘Error and attack tolerance of complex networks’, *Nature*. Nature Publishing Group, 406(6794), pp. 378–382.

- Alva, V., Söding, J. and Lupas, A. N. (2015) ‘A vocabulary of ancient peptides at the origin of folded proteins’, *Elife*. eLife Sciences Publications Limited, 4, p. e09410.
- Apic, G., Gough, J. and Teichmann, S. A. (2001a) ‘An insight into domain combinations’, *Bioinformatics*. Oxford Univ Press, 17(suppl 1), p. S83.
- Apic, G., Gough, J. and Teichmann, S. A. (2001b) ‘Domain combinations in archaeal, eubacterial and eukaryotic proteomes’, *Journal of molecular biology*. Elsevier, 310(2), pp. 311–325.
- Apic, G., Huber, W. and Teichmann, S. A. (2003) ‘Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination’, *Journal of structural and functional genomics*. Springer, 4(2), pp. 67–78.
- Arita, M. (2005) ‘Scale-freeness and biological networks’, *Journal of biochemistry*. Jpn Biochemical Soc, 138(1), p. 1.
- Aroul-Selvam, R., Hubbard, T. and Sasidharan, R. (2004) ‘Domain insertions in protein structures’, *Journal of molecular biology*. Elsevier, 338(4), pp. 633–641.
- Aziz, M. F. *et al.* (2012) ‘Stress induces biphasic-rewiring and modularization patterns in the metabolomic networks of *Escherichia coli*’, *IEEE Intl. Conf. Bioinf. Biomed.*, pp. 593–597. doi: 10.1109/BIBM.2012.6392626.
- Aziz, M. F., Caetano-Anollés, K. and Caetano-Anollés, G. (2016) ‘The early history and emergence of molecular functions and modular scale-free network behavior’, *Scientific reports*. Nature Publishing Group, 6.
- Bajaj, M. and Blundell, T. (1984) ‘Evolution and the tertiary structure of proteins’, *Annual review of biophysics and bioengineering*. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, 13(1), pp. 453–492.
- Barabási, A. L. and Albert, R. (1999) ‘Emergence of scaling in random networks’, *science*. American Association for the Advancement of Science, 286(5439), p. 509.
- Barabási, A. L., Gulbahce, N. and Loscalzo, J. (2011) ‘Network medicine: a network-based approach to human disease’, *Nature Reviews Genetics*. Nature Publishing Group, 12(1), pp. 56–68.
- Barabási, A. L. and Oltvai, Z. N. (2004) ‘Network biology: understanding the cell’s functional organization’, *Nature Reviews Genetics*. Nature Publishing Group, 5(2), pp. 101–113.
- Bashton, M. and Chothia, C. (2002) ‘The geometry of domain combination in proteins¹’, *Journal of molecular biology*. Elsevier, 315(4), pp. 927–939.
- Bashton, M. and Chothia, C. (2007) ‘The generation of new protein functions by the combination of domains’, *Structure*. Elsevier, 15(1), pp. 85–99.
- Berezovsky, I. N., Guarnera, E. and Zheng, Z. (2017) ‘Basic units of protein structure, folding, and function’, *Progress in biophysics and molecular biology*. Elsevier, 128, pp. 85–99.
- Berezovsky, I. N. and Trifonov, E. N. (2001) ‘Van der Waals locks: loop-n-lock structure of globular proteins’, *Journal of molecular biology*. Elsevier, 307(5), pp. 1419–1426.

- Björklund, Å. K. *et al.* (2005) 'Domain rearrangements in protein evolution', *Journal of molecular biology*. Elsevier, 353(4), pp. 911–923.
- Bollobas, B. (1985) *Random Graphs*. Academic Press, London.
- Bonet, J. *et al.* (2014) 'ArchDB 2014: structural classification of loops in proteins', *Nucleic acids research*. Oxford University Press, 42(D1), pp. D315--D319.
- Bornberg-Bauer, E. *et al.* (2005) 'The evolution of domain arrangements in proteins and interaction networks', *Cellular and Molecular Life Sciences CMLS*. Springer, 62(4), pp. 435–445.
- Bornberg-Bauer, E. and Albà, M. (2013) 'Dynamics and adaptive benefits of modular protein evolution', *Current opinion in structural biology*. Elsevier.
- Bornberg-Bauer, E., Huylmans, A.-K. and Sikosek, T. (2010) 'How do new proteins arise?', *Current opinion in structural biology*. Elsevier, 20(3), pp. 390–396.
- Brenner, S. E. and others (2000) 'Target selection for structural genomics', *nature structural biology*. NATURE AMERICA INC, 7(11; SUPP), pp. 967–969.
- Burley, S. K. and others (2000) 'An overview of structural genomics', *Nature Structural Biology*. NATURE AMERICA INC, 7(11; SUPP), pp. 932–934.
- Caetano-Anollés, D. *et al.* (2011) 'Proteome evolution and the metabolic origins of translation and cellular life', *Journal of molecular evolution*. Springer, 72(1), pp. 14–33.
- Caetano-Anollés, G. *et al.* (2009) 'The origin, evolution and structure of the protein world', *Biochem. J*, 417, pp. 621–637.
- Caetano-Anollés, G. *et al.* (2019) 'Emergence of hierarchical modularity in evolving networks uncovered by phylogenomic analysis', *Evolutionary Bioinformatics*. SAGE Publications Sage UK: London, England, 15, p. 1176934319872980.
- Caetano-Anollés, G. and Caetano-Anollés, D. (2003) 'An evolutionarily structured universe of protein architecture', *Genome research*. Cold Spring Harbor Lab, 13(7), pp. 1563–1571.
- Caetano-Anollés, G. and Caetano-Anollés, D. (2005) 'Universal sharing patterns in proteomes and evolution of protein fold architecture and life', *Journal of molecular evolution*. Springer, 60(4), pp. 484–498.
- Caetano-Anollés, G., Kim, H. S. and Mittenthal, J. E. (2007) 'The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture', *Proceedings of the National Academy of Sciences*. National Acad Sciences, 104(22), p. 9358.
- Caetano-Anollés, G., Kim, K. M. and Caetano-Anollés, D. (2012) 'The phylogenomic roots of modern biochemistry: Origins of proteins, cofactors and protein biosynthesis', *Journal of molecular evolution*. Springer, 74(1–2), pp. 1–34.
- Caetano-Anollés, G., Wang, M. and Caetano-Anollés, D. (2013) 'Structural phylogenomics retrodicts the origin of the genetic code and uncovers the evolutionary impact of protein flexibility', *PloS one*. Public Library of Science, 8(8), p. e72225.

- Chothia, C. (1984) ‘Principles that determine the structure of proteins’, *Annual review of biochemistry*. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, 53(1), pp. 537–572.
- Chothia, C. *et al.* (2003) ‘Evolution of the protein repertoire’, *Science*. American Association for the Advancement of Science, 300(5626), p. 1701.
- Chothia, C. and Gough, J. (2009) ‘Genomic and structural aspects of protein evolution’, *Biochem. J*, 419, pp. 15–28.
- Chothia, C., Levitt, M. and Richardson, D. (1977) ‘Structure of proteins: packing of alpha-helices and pleated sheets’, *Proceedings of the National Academy of Sciences*. National Acad Sciences, 74(10), p. 4130.
- Chothia, C. and others (1992) ‘Proteins. One thousand families for the molecular biologist.’, *Nature*, 357(6379), p. 543.
- Cohen, R. and Havlin, S. (2003) ‘Scale-free networks are ultrasmall’, *Physical Review Letters*. APS, 90(5), p. 58701.
- Danos, V. *et al.* (2012) ‘Graphs, rewriting and pathway reconstruction for rule-based models’, in *FSTTCS*, pp. 276–288.
- Debès, C. *et al.* (2013) ‘Evolutionary optimization of protein folding’, *PLoS Comput Biol*. Public Library of Science, 9(1), p. e1002861.
- Deeds, E. J. *et al.* (2012) ‘Combinatorial complexity and compositional drift in protein interaction networks’, *PLoS One*. Public Library of Science, 7(3), p. e32032.
- Deeds, E. J., Bachman, J. A. and Fontana, W. (2012) ‘Optimizing ring assembly reveals the strength of weak interactions’, *Proceedings of the National Academy of Sciences*. National Acad Sciences, 109(7), pp. 2348–2353.
- Dhar, J. and Chakrabarti, P. (2015) ‘Defining the loop structures in proteins based on composite β -turn mimics’, *Protein Engineering, Design and Selection*. Oxford University Press, 28(6), pp. 153–161.
- Dorogovtsev, S. N. and Mendes, J. F. F. (2002) ‘Evolution of networks’, *Advances in physics*. Taylor & Francis, 51(4), pp. 1079–1187.
- Edwards, H., Abeln, S. and Deane, C. M. (2013) ‘Exploring Fold Space Preferences of New-born and Ancient Protein Superfamilies’, *PLoS Computational Biology*, 9(11), p. e1003325. doi: 10.1371/journal.pcbi.1003325.
- Ekman, D. *et al.* (2005) ‘Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions’, *Journal of molecular biology*. Elsevier, 348(1), pp. 231–243.
- Enright, A. J. *et al.* (1999) ‘Protein interaction maps for complete genomes based on gene fusion events’, *Nature*. Nature Publishing Group, 402(6757), pp. 86–90.
- Erdős, P. and Rényi, A. (1960) ‘Connectivity of random nets’, *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5, pp. 17–61.

- Faloutsos, M., Faloutsos, P. and Faloutsos, C. (1999) ‘On power-law relationships of the internet topology’, in *ACM SIGCOMM Computer Communication Review*, pp. 251–262.
- Feller, S. M. and Lewitzky, M. (2012) ‘What’s in a loop?’, *Cell Communication and Signaling*. BioMed Central, 10(1), pp. 1–6.
- Finkelstein, A. V and Ptitsyn, O. B. (1987) ‘Why do globular proteins fit the limited set of folding patterns?’, *Progress in biophysics and molecular biology*, 50(3), p. 171.
- Flaherty, K. M. *et al.* (1991) ‘Similarity of the three-dimensional structures of actin and the ATPase fragment of a 70-kDa heat shock cognate protein’, *Proceedings of the National Academy of Sciences*. National Acad Sciences, 88(11), p. 5041.
- Forslund, K. *et al.* (2008) ‘Domain tree-based analysis of protein architecture evolution’, *Molecular biology and evolution*. SMOE, 25(2), pp. 254–264.
- Fukami-Kobayashi, K. *et al.* (2007) ‘A tree of life based on protein domain organizations’, *Molecular biology and evolution*. SMOE, 24(5), pp. 1181–1189.
- Gerstein, M. (1998) ‘How representative are the known structures of the proteins in a complete genome? A comprehensive structural census’, *Folding and Design*. Elsevier, 3(6), pp. 497–512.
- Goh, K.-I. *et al.* (2007) ‘The human disease network’, *Proceedings of the National Academy of Sciences*. National Acad Sciences, 104(21), pp. 8685–8690.
- Goncarencu, A. and Berezovsky, I. N. (2010) ‘Prototypes of elementary functional loops unravel evolutionary connections between protein functions’, *Bioinformatics*. Oxford Univ Press, 26(18), pp. i497--i503.
- Goncarencu, A. and Berezovsky, I. N. (2012) ‘Exploring the evolution of protein function in Archaea’, *BMC Evolutionary Biology*. BioMed Central Ltd, 12(1), p. 75.
- Goncarencu, A. and Berezovsky, I. N. (2015) ‘Protein function from its emergence to diversity in contemporary proteins’, *Physical biology*. IOP Publishing, 12(4), p. 45002.
- Guharoy, M. and Chakrabarti, P. (2007) ‘Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein--protein interactions’, *Bioinformatics*. Oxford Univ Press, 23(15), pp. 1909–1918.
- Guimera, R., Sales-Pardo, M. and Amaral, L. A. N. (2007) ‘A network-based method for target selection in metabolic networks’, *Bioinformatics*. Oxford Univ Press, 23(13), pp. 1616–1622.
- Huang, H., Jedynek, B. M. and Bader, J. S. (2007) ‘Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps’, *PLoS computational biology*. Public Library of Science, 3(11), p. e214.
- Huss, M. and Holme, P. (2006) ‘Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks’, *Arxiv preprint q-bio/0603038*.
- Jefferson, E. R. *et al.* (2007) ‘SNAPPI-DB: a database and API of structures, interfaces and alignments for protein--protein interactions’, *Nucleic acids research*. Oxford Univ Press, 35(suppl 1), pp. D580--D589.

- Jeltsch, A. (1999) ‘Circular permutations in the molecular evolution of DNA methyltransferases’, *Journal of molecular evolution*. Springer, 49(1), pp. 161–164.
- Jeong, H. *et al.* (2000) ‘The large-scale organization of metabolic networks’, *Nature*. Nature Publishing Group, 407(6804), pp. 651–654.
- John, G. (1990) ‘Six degrees of separation: A Play’. Vintage Books, New York.
- Johnson, S. *et al.* (2010) ‘Entropic origin of disassortativity in complex networks’, *Physical review letters*. APS, 104(10), p. 108702.
- Kashtan, N. and Alon, U. (2005) ‘Spontaneous evolution of modularity and network motifs’, *Proceedings of the National Academy of Sciences*. National Acad Sciences, 102(39), pp. 13773–13778.
- Kashtan, N., Noor, E. and Alon, U. (2007) ‘Varying environments can speed up evolution’, *Proceedings of the National Academy of Sciences*. National Acad Sciences, 104(34), pp. 13711–13716.
- Kim, H., Mittenthal, J. and Caetano-Anollés, G. (2006) ‘MANET: tracing evolution of protein architecture in metabolic networks’, *BMC bioinformatics*. BioMed Central Ltd, 7(1), p. 351.
- Kim, K. M. and Caetano-Anollés, G. (2010) ‘Emergence and evolution of modern molecular functions inferred from phylogenomic analysis of ontological data’, *Molecular biology and evolution*. SMOE, 27(7), pp. 1710–1733.
- Kim, W. K. *et al.* (2006) ‘The many faces of protein--protein interactions: A compendium of interface geometry’, *PLoS computational biology*. Public Library of Science, 2(9), p. e124.
- Kitano, H. (2004) ‘Biological robustness’, *Nature Reviews Genetics*. Nature Publishing Group, 5(11), pp. 826–837.
- Kochen, M. (1989) *The small world*. Ablex Publishing.
- Kummerfeld, S. K. and Teichmann, S. A. (2005) ‘Relative rates of gene fusion and fission in multi-domain proteins’, *Trends in genetics*. Elsevier, 21(1), pp. 25–30.
- Lage, K. *et al.* (2007) ‘A human phenome-interactome network of protein complexes implicated in genetic disorders’, *Nature biotechnology*. Nature Publishing Group, 25(3), pp. 309–316.
- Leszczynski, J. F. and Rose, G. D. (1986) ‘Loops in globular proteins: a novel category of secondary structure’, *Science*. American Association for the Advancement of Science, 234(4778), pp. 849–855.
- Levitt, M., Chothia, C. and others (1976) ‘Structural patterns in globular proteins’, *Nature*, 261(5561), p. 552.
- Levy, E. D. *et al.* (2006) ‘3D complex: a structural classification of protein complexes’, *PLoS computational biology*. Public Library of Science, 2(11), p. e155.
- Li, L. *et al.* (2004) ‘A first-principles approach to understanding the internet’s router-level topology’, in *ACM SIGCOMM Computer Communication Review*, pp. 3–14.

- Li, L. *et al.* (2005) ‘Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications (Extended Version)’, *Arxiv preprint cond-mat/0501169*.
- Liberles, D. A. *et al.* (2012) ‘The interface of protein structure, protein biophysics, and molecular evolution’, *Protein Science*. Wiley Online Library, 21(6), pp. 769–785.
- Lorenz, D. M., Jeng, A. and Deem, M. W. (2011) ‘The emergence of modularity in biological systems’, *Physics of life reviews*. Elsevier, 8(2), pp. 129–160.
- Marcotte, E. M. *et al.* (1999) ‘A census of protein repeats’, *Journal of molecular biology*. Elsevier, 293(1), pp. 151–160.
- Maynard-Smith, J. and Szathmary, E. (1995) ‘The major evolutionary transitions’, *Nature*, 374, pp. 227–232.
- Mengistu, H. *et al.* (2016) ‘The evolutionary origins of hierarchy’, *PLoS computational biology*. Public Library of Science San Francisco, CA USA, 12(6), p. e1004829.
- Milgram, S. (1967) ‘The small world problem’, *Psychology today*. New York, 2(1), pp. 60–67.
- Mittenthal, J. E., Caetano-Anollés, D. and Caetano-Anollés, G. (2012) ‘Biphasic patterns of diversification and the emergence of modules’, *Frontiers in genetics*. Frontiers, 3, p. 147.
- Moore, A. D. *et al.* (2008) ‘Arrangements in the modular evolution of proteins’, *Trends in biochemical sciences*. Elsevier, 33(9), pp. 444–451.
- Moore, A. D. *et al.* (2013) ‘Quantification and functional analysis of modular protein evolution in a dense phylogenetic tree’, *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*. Elsevier.
- Moore, A. D. and Bornberg-Bauer, E. (2012) ‘The dynamics and evolutionary potential of domain loss and emergence’, *Molecular biology and evolution*. SMOE, 29(2), pp. 787–796.
- Mrvar, A. and Batagelj, V. (2016) ‘Analysis and visualization of large networks with program package Pajek’, *Complex Adaptive Systems Modeling*. Springer, 4(1), pp. 1–8. Available at: <http://casmodeling.springeropen.com/articles/10.1186/s40294-016-0017-8>, Date of access: 08/02/2016.
- Murzin, A. G. *et al.* (1995) ‘SCOP: a structural classification of proteins database for the investigation of sequences and structures’, *Journal of molecular biology*. London, New York, Academic Press., 247(4), pp. 536–540.
- Nacher, J. C. *et al.* (2009) ‘A bipartite graph based model of protein domain networks’, *Complex Sciences*. Springer, pp. 525–535.
- Nacher, J. C. and Akutsu, T. (2007) ‘Recent progress on the analysis of power-law features in complex cellular networks’, *Cell biochemistry and biophysics*. Springer, 49(1), pp. 37–47.
- Nacher, J. C., Hayashida, M. and Akutsu, T. (2006) ‘Protein domain networks: Scale-free mixing of positive and negative exponents’, *Physica A: Statistical Mechanics and its Applications*. Elsevier, 367, pp. 538–552.

- Nakashima, H., Nishikawa, K. and Ooi, T. (1986) ‘The folding type of a protein is relevant to the amino acid composition’, *The Journal of Biochemistry*. Oxford University Press, 99(1), pp. 153–162.
- Nasir, A., Kim, K. M. and Caetano-Anollés, G. (2017) ‘Phylogenetic tracings of proteome size support the gradual accretion of protein structural domains and the early origin of viruses from primordial cells’, *Frontiers in microbiology*. Frontiers, 8, p. 1178.
- Nath, N., Mitchell, J. B. O. and Caetano-Anollés, G. (2014) ‘The natural history of biocatalytic mechanisms’, *PLoS Comput Biol*. Public Library of Science, 10(5), p. e1003642.
- Nepomnyachiy, S., Ben-Tal, N. and Kolodny, R. (2014) ‘Global view of the protein universe’, *Proceedings of the National Academy of Sciences*. National Acad Sciences, 111(32), pp. 11691–11696.
- Nepomnyachiy, S., Ben-Tal, N. and Kolodny, R. (2017) ‘Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths’, *Proceedings of the National Academy of Sciences*. National Acad Sciences, 114(44), pp. 11703–11708.
- Newman, M. E. J. (2003) ‘The structure and function of complex networks’, *SIAM review*. JSTOR, pp. 167–256.
- Papaleo, E. *et al.* (2016) ‘The role of protein loops and linkers in conformational dynamics and allostery’, *Chemical reviews*. ACS Publications, 116(11), pp. 6391–6423.
- Pasek, S., Risler, J. L. and Brézellec, P. (2006) ‘Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins’, *Bioinformatics*. Oxford Univ Press, 22(12), pp. 1418–1423.
- Pauling, L., Corey, R. B. and Branson, H. R. (1951) ‘The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain’, *Proceedings of the National Academy of Sciences*. National Acad Sciences, 37(4), pp. 205–211.
- Payne, J. L. and Eppstein, M. J. (2009) ‘Evolutionary dynamics on scale-free interaction networks’, *Evolutionary Computation, IEEE Transactions on*. IEEE, 13(4), pp. 895–912.
- Ravasz, E. *et al.* (2002) ‘Hierarchical organization of modularity in metabolic networks’, *science*. American Association for the Advancement of Science, 297(5586), pp. 1551–1555.
- Richardson, J. S. (1976) ‘Handedness of crossover connections in beta sheets’, *Proceedings of the National Academy of Sciences*. National Acad Sciences, 73(8), p. 2619.
- Richardson, J. S. (1977) ‘Beta-sheet topology and the relatedness of proteins’, *Nature*, 268, pp. 495–500.
- Roy, S. *et al.* (2009) ‘Exploiting amino acid composition for predicting protein-protein interactions’, *PloS one*. Public Library of Science San Francisco, USA, 4(11), p. e7813.
- Rual, J. F. *et al.* (2005) ‘Towards a proteome-scale map of the human protein--protein interaction network’, *Nature*. Nature Publishing Group, 437(7062), pp. 1173–1178.
- Shiu, S. H. *et al.* (2004) ‘Comparative analysis of the receptor-like kinase family in Arabidopsis and rice’, *The Plant Cell Online*. Am Soc Plant Biol, 16(5), pp. 1220–1234.

- Shoemaker, B. A. and Panchenko, A. R. (2007) ‘Deciphering protein--protein interactions. part i. experimental techniques and databases’, *PLoS computational biology*. Public Library of Science, 3(3), p. e42.
- Simon, H. A. (1991) ‘The architecture of complexity’, in *Facets of systems science*. Springer, pp. 457–476.
- Skrabanek, L. *et al.* (2008) ‘Computational prediction of protein--protein interactions’, *Molecular biotechnology*. Springer, 38(1), pp. 1–17.
- Sobolevsky, Y., Guimaraes, R. C. and Trifonov, E. N. (2013) ‘Towards functional repertoire of the earliest proteins’, *Journal of Biomolecular Structure and Dynamics*. Taylor & Francis, 31(11), pp. 1293–1300.
- Söding, J. and Lupas, A. N. (2003) ‘More than the sum of their parts: on the evolution of proteins from peptides’, *Bioessays*. Wiley Online Library, 25(9), pp. 837–846.
- Solé, R. and Valverde, S. (2004) ‘Information theory of complex networks: On evolution and architectural constraints’, *Complex networks*. Springer, pp. 189–207.
- Stelzl, U. *et al.* (2005) ‘A human protein-protein interaction network: a resource for annotating the proteome’, *Cell*. Elsevier, 122(6), pp. 957–968.
- Sternberg, M. J., Thornton, J. M. and others (1976) ‘On the conformation of proteins: the handedness of the beta-strand-alpha-helix-beta-strand unit.’, *Journal of molecular biology*, 105(3), p. 367.
- Tanaka, R. *et al.* (2005) ‘Highly optimised global organisation of metabolic networks’, *Systems biology*. IEE, 152(4), p. 179.
- Tanaka, R. (2005) ‘Scale-rich metabolic networks’, *Physical review letters*. APS, 94(16), p. 168101.
- Tanaka, R. and Doyle, J. (2004) ‘Scale-rich metabolic networks: background and introduction’, *Arxiv preprint q-bio/0410009*.
- Tanaka, R., Yi, T. M. and Doyle, J. (2005) ‘Some protein interaction data do not exhibit power law statistics’, *FEBS letters*. Elsevier, 579(23), pp. 5140–5144.
- Taylor, W. R. (2007) ‘Evolutionary transitions in protein fold space’, *Current opinion in structural biology*. Elsevier, 17(3), pp. 354–361.
- Teichmann, S. A., Chothia, C. and Gerstein, M. (1999) ‘Advances in structural genomics’, *Current opinion in structural biology*. Elsevier, 9(3), pp. 390–399.
- Teichmann, S. A., Park, J. and Chothia, C. (1998) ‘Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements’, *Proceedings of the National Academy of Sciences*. National Acad Sciences, 95(25), p. 14658.
- Teyra, J. *et al.* (2008) ‘SCOWLP classification: structural comparison and analysis of protein binding regions’, *BMC bioinformatics*. BioMed Central Ltd, 9(1), p. 9.

- Tóth-Petróczy, Á. and Tawfik, D. S. (2014) 'The robustness and innovability of protein folds', *Current opinion in structural biology*. Elsevier, 26, pp. 131–138.
- Trifonov, E. N. and Frenkel, Z. M. (2009) 'Evolution of protein modularity', *Current opinion in structural biology*. Elsevier, 19(3), pp. 335–340.
- Tuncbag, N. *et al.* (2008) 'Architectures and functional coverage of protein-protein interfaces', *Journal of molecular biology*. Elsevier, 381(3), pp. 785–802.
- Vogel, C., Bashton, M., *et al.* (2004) 'Structure, function and evolution of multidomain proteins', *Current opinion in structural biology*. Elsevier, 14(2), pp. 208–216.
- Vogel, C., Berzuini, C., *et al.* (2004) 'Supra-domains: evolutionary units larger than single protein domains', *Journal of molecular biology*. Elsevier, 336(3), pp. 809–823.
- Vogel, C., Teichmann, S. A. and Pereira-Leal, J. (2005) 'The relationship between domain duplication and recombination', *Journal of molecular biology*. Elsevier, 346(1), pp. 355–365.
- Wang, M. *et al.* (2006) 'A phylogenomic reconstruction of the protein world based on a genomic census of protein fold architecture', *Complexity*. Wiley Online Library, 12(1), pp. 27–40.
- Wang, M. *et al.* (2007) 'Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world', *Genome research*. Cold Spring Harbor Lab, 17(11), pp. 1572–1585.
- Wang, M. *et al.* (2011) 'A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation', *Molecular biology and evolution*. SMOE, 28(1), pp. 567–582.
- Wang, M. and Caetano-Anollés, G. (2006a) 'Evolution inferred from domain combination in proteins', *Mol. Biol. Evol*, 23, pp. 2444–2454.
- Wang, M. and Caetano-Anollés, G. (2006b) 'Global phylogeny determined by the combination of protein domains in proteomes', *Molecular biology and evolution*. SMOE, 23(12), pp. 2444–2454.
- Wang, M. and Caetano-Anollés, G. (2009) 'The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world', *Structure*. Elsevier, 17(1), pp. 66–78.
- Watts, D. J. (2004) 'The "new" science of networks', *Annual review of sociology*. JSTOR, pp. 243–270.
- Watts, D. J. and Strogatz, S. H. (1998) 'Collective dynamics of 'small-world' networks', *Nature*. Nature Publishing Group, 393(6684), pp. 440–442.
- Winter, C. *et al.* (2006) 'SCOPPI: a structural classification of protein--protein interfaces', *Nucleic Acids Research*. Oxford Univ Press, 34(suppl 1), pp. D310--D314.
- Wolf, Y. I., Grishin, N. V and Koonin, E. V (2000) 'Estimating the number of protein folds and families from complete genome data', *Journal of Molecular Biology*. Elsevier, 299(4), pp. 897–905.
- Wu, X. *et al.* (2008) 'Network-based global inference of human disease genes', *Molecular Systems Biology*. Nature Publishing Group, 4(1).

Wuchty, S. (2001) 'Scale-free behavior in protein domain networks', *Molecular Biology and Evolution*, 18(9), pp. 1694–1702. doi: 10.1093/oxfordjournals.molbev.a003957.

Wuchty, S. (2006) 'Topology and weights in a protein domain interaction network--a novel way to predict protein interactions', *BMC genomics*. BioMed Central Ltd, 7(1), p. 122.

Yanai, I., Derti, A. and DeLisi, C. (2001) 'Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes', *Proceedings of the National Academy of Sciences*. National Acad Sciences, 98(14), p. 7940.

Yang, S., Doolittle, R. F. and Bourne, P. E. (2005) 'Phylogeny determined by protein domain content', *Proceedings of the National Academy of Sciences of the United States of America*. National Acad Sciences, 102(2), p. 373.

CHAPTER 2: EVOLUTION OF NETWORKS OF PROTEIN DOMAIN ORGANIZATION

2.1 Abstract

Domains are the structural, functional and evolutionary units of proteins. They combine to form multidomain proteins. The evolutionary history of this molecular combinatorics has been studied with phylogenomic methods. Here, we construct networks of domain organization and explore their evolution. These time-varying (evolving) networks revealed two ancient waves of structural novelty arising from ancient ‘*p*-loop’ and ‘winged helix’ domains and a massive ‘big bang’ of domain organization. The evolutionary recruitment of domains was highly modular, hierarchical and ongoing. Domain rearrangements elicited non-random and scale-free network structure. Comparative analyses of preferential attachment, randomness and modularity of networks showed yin-and-yang complementary transition patterns along the evolutionary timeline. Remarkably, evolving networks highlighted a central evolutionary role of cofactor-supporting structures of non-ribosomal peptide synthesis (NRPS) pathways, likely crucial to the early development of the genetic code. Some highly modular domains featured dual response regulation in two-component signal transduction systems with DNA-binding activity linked to transcriptional regulation of responses to environmental change. Interestingly, hub domains across the evolving networks shared the historical role of DNA binding and editing, an ancient protein function in molecular evolution. Our investigation unfolds historical source-sink patterns of evolutionary recruitment that further our understanding of protein architectures and functions.

Index terms: Domains, multidomains, evolution, time events, age, network, connectivity, modularity, randomness, scale-free, scale-rich.

2.2 Introduction

The biological functions of genes manifest through the proteins or functional RNA molecules they encode. In evolution, novel functions appear when genes produce new genes by duplication, mutation, recombination, fusion and fission, or when genes are generated *de novo*. Research has attempted to quantitatively describe the origins of these processes of molecular diversification and how they increase molecular complexity over the course of evolution, for instance through pathways of protein domain organization(Chothia *et al.*, 2003; Chothia and Gough, 2009). Protein domains are structural and functional units of evolution that make up proteins(Wetlaufer, 1973; Janin and Wodak, 1983; Han *et al.*, 2007), sometimes in unusually complex arrangements(Apic, Gough and Teichmann, 2001b)(Wang and Caetano-Anollés, 2009). They fold into compact 3-dimensional (3D) atomic structures that arrange alpha-helical and beta-sheet structure elements into tightly packed conformations of the polypeptide chain(Gustavo Caetano-Anollés, Wang, *et al.*, 2009). The Structural Classification of Proteins (SCOP)(Murzin *et al.*, 1995) and its extended version SCOPe(Chandonia, Fox and Brenner, 2017) are popular taxonomy gold standards of domain structure. SCOP definitions can be used to scan genome sequences for motifs of domains and study how they combine in proteins(Apic, Gough and Teichmann, 2001b). In SCOP, the structure of domains exhibiting similar 3D arrangements of secondary structures and thus identical topologies have been classified as folds (F)(Murzin *et al.*, 1995). Within folds, protein domains whose structure and functional features indicate a common evolutionary origin are further grouped into fold superfamilies (FSF). These FSFs sometimes hold multiple evolutionarily related families, which unify domains with pairwise amino acid identities of more than 30% (Supplementary Fig. S1A). As of March 9, 2021, 276,231 annotated SCOPe domains populate the 175,282 protein structures of the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB-

PDB). We note that the cornerstone of the SCOP domain hierarchy is common ancestry, i.e. the existence of shared-and-derived features in domain sequence, structure and function. Homology is also central to many other domain taxonomies, including CATH(Dawson *et al.*, 2017), CDD(Marchler-Bauer *et al.*, 2015), ProDom(Bru *et al.*, 2005), Pfam(Finn *et al.*, 2008), and the meta-database InterPro(Finn, Attwood and Babbitt, 2017). Most databases benefit from machine learning. For example, SCOP and Pfam depend on the identification of conserved regions in protein sequences through sequence alignment and background knowledge, which are then used to build probabilistic hidden Markov models (HMMs) of linear sequence analysis. For example, SCOP uses HMMs of structural recognition to recurrently enrich the database(Wilson *et al.*, 2009) in a framework that increases alignment-quality and stability of family and superfamily relationships. A similar framework drives the Pfam database but focuses exclusively on sequence information. One difficulty is that not all domains fold into discrete structural entities within the space of possible folds(Harrison *et al.*, 2002). Some popular domains overlap within a continuum. This ‘gregariousness’ complicates domain classification, demanding the exploration of super-secondary structural motifs as candidate lower-level units of structure, function and evolution(Berezovsky, Guarnera and Zheng, 2017).

Domain structures appear repeatedly in the protein molecules, singly or in combination with other domains(Wang and Caetano-Anollés, 2009). More than two-thirds of protein sequences are longer than an average domain length, a vast majority of which are multidomain proteins(M Gerstein, 1998). A study of protein structures in 749 genomes showed that the lengths of orthologous protein families in Eukarya were almost double the lengths found in Bacteria and Archaea(Wang, Kurland and Caetano-Anollés, 2011). This variance among lengths results from shorter prokaryotic nondomain sequences that link domains to each other in proteins and have evolved reductively in

prokaryotes but not in eukaryotes. The arrangement of domains along the sequence of multimeric proteins is referred to as '*domain organization*.' Both the structure and organization of domains, which have been collectively termed protein domain '*architecture*', are considered far more evolutionarily conserved than protein sequence(Chothia and Lesk, 1986; Bashton and Chothia, 2002; Illergård, Ardell and Elofsson, 2009; Wang and Caetano-Anollés, 2009). In addition, some domain combinations make up functional units that recur in different protein contexts(Vogel, Berzuini, *et al.*, 2004). They have been termed supradomains (Supplementary Fig. S1B). Thus, domains and supradomains behave as modules, parts that interact more often with each other than with other parts or modules of the system.

Comparative genomic approaches allow to study the modular landscape of domain organization. For example, the evolutionary placement of domains in multiple architectural contexts can be quantified by counting distinct neighbors(Wuchty, 2001), domain adjacencies(Apic, Huber and Teichmann, 2003), or consecutive domain triplets(Tordai *et al.*, 2005) in proteins. These measures of 'versatility', 'promiscuity' or 'mobility' (reuse) of domain building blocks depend on both domain size and abundance. Smaller domains are more likely to be used in multidomain proteins and are therefore more mobile(Tordai *et al.*, 2005; Weiner, Moore and Bornberg-Bauer, 2008), an observation supported by a Menzerath-Altmann's law of domain organization driven by an economy of scale(Shahzad, Mittenthal and Caetano-Anollés, 2015). Similarly, highly abundant domains appear more versatile, prompting abundance-based normalization of domain versatility measurements when studying intrinsic combinatorial properties and variation across lineages and biological functions(Christine Vogel, Teichmann and Pereira-Leal, 2005; Basu *et al.*, 2008; Xie, Jin and Mao, 2011).

In order to retrace past events in architectural evolution, statements of history (phylogenies) proposed directly from genomic data must be used to build chronologies of first evolutionary appearance of domains and domain architectures. Unfortunately, protein sequence has limited power in deep retrodictive exploration(Chothia and Lesk, 1986; Bashton and Chothia, 2002; Illergård, Ardell and Elofsson, 2009; Wang and Caetano-Anollés, 2009). Furthermore, while structure is conserved over longer evolutionary timescales, a general metric for global pairwise comparison of structures does not yet exist(William R. Taylor, 2007). Thus, the systematic classification of protein structure has been unable to unify the widely divergent folded structures at any level of abstraction (e.g. FSFs in a ‘galaxy’ of folds(Alva *et al.*, 2010)), likely because different neighborhoods in protein sequence space contain different structures and functions(Ferrada and Wagner, 2010). Construction of a ‘periodic table’ of idealized structural representations of folds(Taylor, 2002) has not alleviated this difficulty due to an absence of rules of structural transformation that would explain the comparative framework. Efforts to dissect the evolution of domain architectures has recently been reviewed(Forslund and Sonnhammer, 2012). To overcome limitations and produce global evolutionary views of the protein world, the focus must shift from molecules to proteomes. Trees of proteomes were first reconstructed from a proteomic census of structural domains (beginning with ref.(Mark Gerstein, 1998)), and were later used to trace character-state changes along their branches to establish possible domain origins(Abeln and Deane, 2005; Edwards, Abeln and Deane, 2013). This approach, however, was restricted to domain structures and architectures appearing after the common ancestor of the proteomes surveyed in the trees. A much more effective way to create truly global chronologies of the protein world was the reconstruction of phylogenomic trees of domain organization (beginning with ref.(Caetano-Anollés and Caetano-Anollés, 2003)). These phylogenies take advantage of

powerful serial homologies defined by the proteomic abundance of domains(Wang *et al.*, 2007) or architectures(Wang and Caetano-Anollés, 2006b, 2009) defined at F and FSF levels. Phylogenomic trees of domain structures helped uncover the natural history of biocatalysis by tracing chemical mechanisms in enzymatic reactions(Nath, Mitchell and Caetano-Anollés, 2014), analyze the optimization and increase of protein folding speed derived from a flexibility-correlated factor known as contact order (the average relative distance of amino acid contacts in the tertiary structure of proteins)(Debès *et al.*, 2013), and study the history of an ‘elementary functionome’ with a bipartite network of elementary functional loop sequences and structural domains of proteins(Aziz, Caetano-Anollés and Caetano-Anollés, 2016). This last study revealed two initial waves of functional innovation involving founder ‘*p*-loop’ and ‘winged helix’ domain structures, and the emergence of hierarchical modularity and power law behavior in network evolution. Phylogenomic trees of domain architectures and their associated chronologies of molecular accretion showed that architectural diversification evolved through gradual accumulation of domains (singly occurring domains), domain pairs (two different domains), multidomains (numerous domains, with occasional repetition) and domain repeats (domains of one type that are repeated)(Wang and Caetano-Anollés, 2009). The diversification began with a few single-domain architectures earlier in the timeline, followed by an increasing rate of accretion that culminated in a massive “big bang” of domain organization. The accumulation of architectures continued to date but with a decreasing rate(Wang and Caetano-Anollés, 2009; Aziz, Caetano-Anollés and Caetano-Anollés, 2016).

Here, we explore the evolving interactome of protein domain organization. We generate a chronology that captures the historical development of domain and multidomain interactions with a graph theoretical approach(Apic, Gough and Teichmann, 2001b) of time-varying (evolving)

network structure. A timeline rendering the chronology was calibrated with a molecular clock of protein structures, which assigns relative ages of the origin of domain organization to billions of years (Gy) of geological time(Wang *et al.*, 2011). Five distinct composition- and topology-based ‘operative’ criteria of connectivity defined nodes and links of the evolving networks. This strategy identified connectivity distributions in a series of 169 growing networks, hubs of evolutionary recruitment acting as *donors* and *acceptors*, and structural adaptations of evolving networks to modular, random and scale-free properties. In particular, we discover a pattern of connectivity driven by fusions and fissions, respectively, with densely linked older and younger architectures from the evolutionary timeline sandwiching a period of sparse connectivity. This supports a biphasic or hourglass pattern previously observed in protein evolution(Caetano-Anollés *et al.*, 2011) and follows a model of module emergence(Mittenthal, Caetano-Anollés and Caetano-Anollés, 2012). We thus reveal remarkable patterns of emergence of hierarchy, modularity and structural cooption in evolving networks.

2.3 Results and Discussion

2.3.1 Construction of evolving networks

We build a time series of networks of domain organization embedding evolutionary information derived from the sequence and structure of millions of protein sequences encoded in hundreds of genomes. The goal is to unfold the to date history of how single-domain and multidomain proteins share domain make-up and how recruitment processes shape protein evolution. An ‘entity set’ of domains, supradomains, and multidomains were first extracted from the genomic census of fold structure and domain organization. This set of component parts of proteins, mostly recurrent, defined the nodes of the networks, which were labeled with *concise classification strings (ccs)* describing SCOP domain constituents (Fig. 2.1A). We define supradomains as sub-combinations

of domains that appear in the census and are often used as evolutionary building blocks of multidomains. The definition is more inclusive than that of ref(Vogel, Berzuini, *et al.*, 2004).

The growing interactions among contemporary architectures are constrained by domain make up and domain arrangement in the protein chain. These evolving interactions were captured with five different operative criteria for timed network generation defined by composition, pairwise occurrence, adjacency, and splicing of domain parts in a protein molecule, where: (i) *composition* describes makeup (component parts) of the molecular whole; (ii) *pairwise occurrence* describes appearance of parts in sets of two; (iii) *adjacency* refers to their geometrical or spatial arrangement (topology); and (iv) *splicing* refers to the rearrangement of parts by operations of joining and excision that decompose structures (Fig. 2.1B). The Composition Network (CX) linked domain and supradomain to multidomain nodes (in a partially bimodal fashion) when proteins shared compositional makeup. The Pairwise Network (PX) connected domain to supradomain nodes when components occurred in pairs in a protein. The Pairwise Adjacency Network (PAX) connected domain to supradomain nodes when components occurred in pairs that were adjacent. The Spliced Pairwise Network (SPX) linked domain nodes to each other when their pairs were present in domain-spliced proteins. Lastly, the Spliced Pairwise Adjacency Network (SPAX) linked domain nodes to each other when their adjacent pairs were present in the domain-spliced proteins (Fig. 2.2).

We then mapped the evolutionary ages of architectures onto the nodes of networks built using these five operative criteria (Supplementary Fig. S2). We did so for each of the 169 time-events of the timeline. Network construction has been illustrated with connectivity details of the most ancient domains (Supplementary Fig. S3) and further described in *Section 1* of Supplementary Text. Networks showcased time directionality, connectivity distributions, and network layouts:

(i) *Time Directionality*: Mapping ages onto networks helped follow their evolutionary growth, as nodes and links accumulated over time since the origin of proteins to the present. The timeline of networks imposed a time directionality on network links, making them *arcs* (directed edges with arrows pointing from older to younger nodes) of directed graphs (Fig. 2.1C). The ages of arcs were borrowed from the youngest of the component nodes involved in a link (Supplementary Fig. S3B).

(ii) *Degree Distributions*: The number of links connected to a node define that node's 'degree'. The degree distribution is a 'composability' attribute of a network and the entity set represented by its nodes, a design principle describing the inter-relationship of components of a system. In network evolution, the appearance of a new node may trigger establishment of one or more arcs from existing (older) nodes. Furthermore, *outdegree* describes the number of outward links and *indegree* the number of inward links from a node. As the timeline progresses, older nodes gain higher outdegrees as compared to the higher indegrees of recent nodes (Fig. 2.1C), polarizing the network with arcs depicting 'arrows of time' (Supplementary Figs. S2 and S3). The chronological appearance of architectures (domains, supradomains and multidomains) as network connectivity expands along the timeline causes degree to accumulate in the evolving networks (Fig. 2.2). Multiple interactions of nodes along the timeline diversified connectivity, a feature captured and quantified by weighted degree. Interestingly, box-and-whisker's plots of weighted outdegree and indegree demonstrate bimodal degree distributions typical of biological systems (Aziz *et al.*, 2012; Mittenthal, Caetano-Anollés and Caetano-Anollés, 2012) (Supplementary Fig. S4). The yin-yang patterns of contractions and expansions of architectural innovation are evident from the distributions of modern outdegrees and indegrees (Supplementary Fig. S5). In particular, the cumulative outdegree and indegree scattergrams demonstrate an hourglass (or bimodal) pattern of linkage development unfolding in evolution (Supplementary Fig. S6).

(iii) *Time Event-based ‘Radial’ and ‘Waterfall’ Layouts*: The growth of a network evolving at discrete temporal intervals can be modeled with Discrete Event Simulation (DES) tools (MacDougall, 1987; Delaney and Vaccari, 1989; Pidd, 2006). Borrowing the DES rationale, we modeled the evolution of directed networks of domain organization with time flowing from one event to another as discrete evolutionary ‘time steps’, typical of a *step function*. The progression of events was visualized with two types of layouts, a vertical representation we coined ‘*waterfall*’ layout that had nodes arranged top-down by age and a concentric ‘*radial*’ representation of growing networks that unfolded time-events of protein evolution from center to periphery (Fig. 2.1C). Network clusters comprising of hubs and their cohesive neighbors were segregated to improve differentiation along the horizontal axis. The *waterfall* and *radial* layouts made evolutionary recruitment evident as time events progressed downward or outward, respectively (Figs. 2.2 and 2.3).

2.3.2 Early history of modern domain organization

The accumulation of links connecting domain, supradomain and multidomain proteins in evolving CX, PX, PAX, SPX and SPAX networks played back the complicated history of domain recruitments that drive the evolution of domain organization. Figure 2.2 shows networks in radial layout at representative time-events defining boundaries of the three epochs of the evolving protein world (‘architectural diversification’, ‘superkingdom specification’ and ‘organismal diversification’, *sensu* (Wang *et al.*, 2007; Wang and Caetano-Anollés, 2009)). Networks grew in time and became increasingly complicated tangles, massively expanding after a “big bang” of domain combinations during the organismal diversification epoch. Movies described the evolutionary growth of these networks (Supplementary Video 1).

To illustrate the versatility of the waterfall visualization strategy, we dissected the early origin of proteins with the SPX network. Two major waves of structural innovation arising from ancient ‘*p*-loop’ and ‘winged helix’ domains were observed in the waterfall diagrams of a highly connected (reduced) subnetwork visualization of the SPX network (Fig. 2.3), matching similar recruitment waves observed in the study of evolutionary networks of elementary functionomes (Aziz, Caetano-Anollés and Caetano-Anollés, 2016) and metabolites (Caetano-Anollés, Kim and Mittenthal, 2007). Waves originated in primordial $\alpha/\beta/\alpha$ -layered sandwich, β -barrel and helical bundle structures identified in an earlier structural phylogenomic study as part of the most ancient 54 protein domain families (Caetano-Anollés, Kim and Caetano-Anollés, 2012). However, most of the connectivity of these major pathways was established during the organismal diversification epoch less than 1.5 Gy ago ($nd \geq 0.6$) and hence was fully developed relatively recently in evolution. The ‘*p*-loop’ and ‘winged helix’ waves embedded the major gateways of enzymatic recruitment we previously reported for metabolism (Caetano-Anollés, Kim and Mittenthal, 2007). The first gateway was mediated by the c.37 P-loop hydrolase fold and originated in the energy interconversion pathways of the purine metabolism subnetwork. The second pathway was mediated by the a.4 winged helix fold and originated in the biosynthesis of cofactors and the metabolic subnetwork of porphyrin and chlorophyll (Caetano-Anollés, Kim and Mittenthal, 2007; Gustavo Caetano-Anollés, Yafremava, *et al.*, 2009; Caetano-Anollés and Caetano-Anollés, 2013). The congruent realization of these evolutionary patterns with data sources of different types is remarkable (Supplementary Video 2). It strongly supports the historical statements we propose. Further information can be found in *Section 2* of Supplementary Text.

2.3.3 Network analysis of cooption mechanisms of recruitment

The networks of domains (SPX and SPAX) elicited 161 unique time-events along the evolutionary timeline, out of a total 169 events expected for networks of domains, supradomains and domain combinations (CX, PX and PAX) (Supplementary Tables 1-5). The node and connectivity distributions among the time-event bins of the evolving networks highlight the widespread, growing and recurrent combinatorial recruitment process that incorporates domains and their combinations into protein scaffolds and drives structural evolution (Fig. 2.2). Indeed, the largest hubs representing the most popular domains in the highly connected SPX subnetwork appeared not only early in evolution but also in the modern protein world (Fig. 2.3). Similar to the evolution of elementary functions (Aziz, Caetano-Anollés and Caetano-Anollés, 2016), domain innovation also developed early during the first ~1.8 Gy of protein history (Fig. 2.3). The combinatorial recruitment process however spanned the entire timeline (Supplementary Fig. S2). In terms of origins, the first donor and acceptor composition event occurred in protein evolution with the appearance of a link in the CX network connecting domain c.2.1 to domain combination c.2.1|a.100.1, ~3.54 Gya ($nd = 0.069$). The first donor and acceptor pair occurred in the pairwise PX and SPX networks ~3.12 Gya ($nd = 0.179$), ~0.42 Gy later ($\Delta nd = 0.11$). The pairing event involved domains c.37.1 and d.14.1. The first adjacent donor and acceptor pair of the adjacency-based PAX and SPAX networks appeared ~2.90 Gya ($nd = 0.237$), ~0.22 Gy later ($\Delta nd = 0.06$). The adjacently paired nodes were domains c.37.1 and c.23.16. These observations highlight a remarkable tendency of domain organization to gradually but recurrently constrain pairwise occurrences in multidomain proteins. The evolutionary history of donors and acceptors of domain organization is hence associated with a highly optimized process of cooption. To explore this combinatorics, first we dissected the network connectivity with bar plots that describe the

chronological accumulation of links along the evolutionary timeline (Supplementary Fig. S7). This made general patterns quantitative and source-sink relationships explicit. Second, we analyzed the per unit donor/acceptor ratio in the evolving networks to highlight pairwise cooption and composability, respectively (Supplementary Fig. S8). Specifically, domain acceptors (represented by network indegree) of SPX increased in number to a global average of $8.63 (\pm 0.15)$ sinks per domain in evolution. Domain donors (represented by network outdegree) of SPX reached a higher global average of $9.7 (\pm 0.56)$ sources per domain, indicating significant reutilization of relatively ancient domains. In contrast, the average number of donors and acceptors in the evolving CX network plateaued at 3.41 ± 0.34 sources and 3.43 ± 0.05 sinks per domain/multidomain, respectively. This showed uniform source/sink evolutionary rates as proteins acquired higher composability with time. Third, an inferential analysis of cooption-based source-sink relationships maturing at modern times revealed an independence of patterns from the selected network generation criteria (Supplementary Fig. S9). Primarily, the composition events yielding source domains and supradomains were dominant, with the number of events almost doubling in the CX network from the origin to the organismal diversification epoch ~ 1.5 Gya ($nd = 0.6$). However, the pairwise cooption events of the SPX domain network, e.g., doubled in number and reached relatively comparable levels in evolution only after delays of ~ 0.6 Gy ($\Delta nd = 0.15$) and ~ 2.1 Gya ($nd = 0.75$), respectively. Moreover, the number of cooption events yielding sink domains in SPX almost tripled by the beginning of the organismal diversification epoch. In contrast, the number of CX sinks reached that level only halfway along that evolutionary epoch. These divergent patterns indicate a frustrated dynamics of network growth. The early adoption of composability of domains and supradomains in multidomains seems to have preceded the pairwise cooption of domains in protein history, leading to the numerous recruitment pathways of the modern protein world. A

discussion on the source-sink relationships impacted by domain fusion and fission processes can be found in *Section 3* of Supplementary Text.

2.3.4 Hubs in network evolution

Network hubs are at the heart of network connectivity and could chaperone network evolution (Apic, Huber and Teichmann, 2003). We ranked modern domains and domain combinations of age $nd = 1$ as hubs based on the 99.9th percentile of indegree and outdegree. Hubs were annotated with domain organization attributes, including SCOP domain descriptions, age, fusional/fissional information, and GO terms. We also associated hubs with age ranks reflecting their order of evolutionary appearance in the timeline.

The most notable donor hubs for all networks types were the carrier protein domains e.23.1, a.28.1 and c.69.1, which are involved in Non-Ribosomal Peptide Synthesis (NRPS), whether directly or indirectly through other pathways (Table 2.1). These domains diversified later in evolution yielding cofactor-binding molecular switches and barrel structures (Caetano-Anollés, Kim and Caetano-Anollés, 2012). Ancient NRPS pathways of domain accretion have been associated with a model that not only described stabilization and decoration of membranes by primordial alpha-helical bundles and beta-sheets, but also explained primordial protein synthesis and genetic code specificity chaperoned by ancient forms of aminoacyl-tRNA synthetase (aaRS) catalytic domains and NRPS modules. NRPS even preceded the emergence of the ribosome, acting as scaffold for nucleic acids and the modern translation function. In particular, the PX and PAX networks highlight the central evolutionary role of these novel emerging cofactor structures in the NRPS pathways. Thus, our findings made explicit that our connectivity criteria of generating networks of domain organization were at the cornerstone of the early development of genetic code and

supported the evolutionary model of early biochemistry based on phylogenomic information and network structure.

Domains c.30.1, b.1.1, d.142.1 and g.3.11 ($0.723 < nd < 0.977$) were the most prominent acceptor hubs (Table 2.2). These structures are integral parts of two-component signal transduction systems that are common in microbes. The highly modular domains feature dual response regulator proteins involved in the two-component signal transduction system comprising of an N-terminal response regulator receiver domain and a variable C-terminal effector domain with DNA-binding activity. These proteins are transcriptional regulators in bacteria and some protozoa, detecting and responding to environmental changes, e.g. nitrogen fixation. These evolving interactions of microbes with the environment mediated by two-component systems have apparently influenced the evolutionary process of cooption. Three acceptor hubs that were significant in PX with indegree > 250 (following behind the 99.9th percentile in other networks) were Nucleotide cyclase (d.58.29), Spermadhesin, CUB domain (b.23.1), and Fibronectin type III (b.1.2) ($nd = 0.723$ - 0.809). See *Section 4* of Supplementary Text for additional donor/acceptor hub information, and *Section 5* for cooption events occurring during the ‘big bang’ of domain organization.

2.3.5 Emergence of preferential attachment in network evolution

Genomic-centric processes such as duplication, recombination, fusion and fission shape patterns of molecular complexity(Chothia *et al.*, 2003). Many of these patterns can be explained with large ‘scale-free’ networks that grow by following the preferential attachment principle(Barabási and Albert, 1999). These self-organizing and highly inhomogeneous networks attach links to highly connected hub-like nodes in a ‘rich-get-richer’ fashion, lacking a characteristic scale, irrespective of the properties of individual nodes or systems(Pang and Maslov, 2013). This pattern of network

expansion, which is remarkably popular in biology(Ravasz *et al.*, 2002), is sharply distinct from that of the Erdős–Rényi random network model(Erdős and Rényi, 1960; Bollobas, 1985). In a scale-free network, the probability $P(k)$ of nodes connecting with neighboring k nodes (i.e. the ratio of nodes with k links) decays as a power law, $P(k) \sim k^{-\gamma}$, with γ defined as the exponent of power law decay. The frequency distributions of node-connectivity in biomolecular networks have γ typically ranging 2.1–2.4(Strogatz, 2001). Thus, scale-free properties drive degree distributions entailing heavy tails, where very few nodes have high degree values.

Our statistical analyses of the featured indegree distributions along the timeline of growing networks uncovered interesting patterns of power law dynamics (Fig. 2.4). The scale-free patterns were established early on in protein evolution, primarily evident in the CX composition network. These patterns were remarkably divergent from evolving networks connected at random (RVN p -value > 0.05). While power law behavior generally declined as the networks evolved (KS p -value < 0.05, α < 2.5), it somewhat sustained after the ‘big bang’ but only in CX and not in the pairwise networks (KS fit and γ closer to 0 and 2 in CX, respectively). A $\log$ linear regression model of CX produced the highest absolute value for γ of 3.81 among the five networks, which was achieved early along the evolutionary timeline ($nd \sim 0.25$). This value of γ was much higher than values reported for metabolic networks ($\gamma \sim 2.2$)(Ravasz *et al.*, 2002). Remarkably, the γ was maintained at ~ 3 before and after the ‘big bang’, while remaining at ~ 2 until modern times with a minimum value of 1.7. The other four networks generated primarily with pairwise criterion apparently deviated from the power-law behavior, especially after the ‘big bang’. For instance, the γ of PX and PAX peaked at 2.4 ($nd \sim 0.35$) and 3.2 ($nd \sim 0.38$), respectively, slightly later than CX. We also noted a transition in γ from 2.1 in PX and 2.7 in PAX prior to the ‘big bang’ to 1.6 in both after the big bang, plateauing at ~ 1 until the present. In the SPX and SPAX networks, γ reached a

peak even later in time than PX and PAX with values of 2.8 ($nd \sim 0.54$) and 3.4 ($nd \sim 0.66$), respectively. These values transitioned from 2.4 in SPX and 2.8 in SPAX from before the big bang to 1.6 and 1.7 after the big bang, respectively, plateauing at ~ 1 in both the networks. As expected, the average γ based on less representative outdegree of each of the five networks remained low (1 ± 0.05).

We noticed that the patterns of $-\gamma$ curves over the connectivity of the networks were biphasic, with two minima at $nd \sim 0.37$ and ~ 0.67 . Moreover, the scale-free tendency of adjacency networks seemed comparatively higher than that of networks lacking the adjacency restriction. For instance, the average values of γ for the PAX and SPAX networks (1.87 ± 0.06 and 2.13 ± 0.07 , respectively) were relatively higher than those for the corresponding parent PX and SPX networks (1.61 ± 0.05 and 1.89 ± 0.06 , respectively). This suggests that the proximity of residuals in the amino acid sequence plays a major role in rendering the power-law behavior of evolving networks of domain organization. Overall, the average γ of CX (2.56 ± 0.06) remained the highest along the evolutionary timeline, indicating that composition strongly elicits the preferential attachment property. A complementary transition from random to non-random behavior (RVN p-value: $1 \rightarrow 0$) in ancient networks ($nd \sim 0.3$) implies deviation from randomness as biological networks evolve. Remarkably, this transition event coincides with the origin of a processive ribosome. Such biphasic patterns are common in biology and have explained the emergence of biological modules (Mittenthal, Caetano-Anollés and Caetano-Anollés, 2012) in metabolic networks of *Escherichia coli* (Aziz *et al.*, 2012), networks of elementary functions (Aziz, Caetano-Anollés and Caetano-Anollés, 2016), and molecular ancestry networks of enzymes (Mughal and Caetano-Anollés, 2019). Section 6 of Supplementary Text further discusses *scale-freeness* and *randomness* of networks.

2.3.6 Emergence of hierarchical modularity

Modular networks embed sets of communities (closely-knit modules) that establish links preferentially within themselves and do so sparsely with the rest (Newman and Girvan, 2004). Network modularity usually offsets the power-law behavior of biological networks by distributing node degrees within communities (Jeong *et al.*, 2000; Overbeek *et al.*, 2000; Wagner and Fell, 2001). However, both scale-free properties and modular structure may co-exist in a network when modules coalesce hierarchically (Ravasz *et al.*, 2002). A primary index of modularity is the *average clustering coefficient* (C), defined as a node-averaged ratio of triangles (graph cycles of length 3) to triads (the connected graph triples) of the network, not taking into account the weights or direction of the node-links (Wasserman and Faust, 1994; Ravasz *et al.*, 2002; Barrat *et al.*, 2004) (Fig. 2.5). The adjacency PAX and SPAX networks both showed the lowest C (averaged over nd) with a value of 0.09 ± 0.009 . The composition CX network had a relatively higher C of 0.2 ± 0.009 . However, the non-adjacency pairwise PX and SPX networks had the highest C values of 0.5 ± 0.02 and 0.32 ± 0.014 , respectively. These values were still lower than those reported for metabolic networks ($C = \sim 0.6$) (Watts and Strogatz, 1998; Wagner and Fell, 2001; Ravasz *et al.*, 2002). Hence, the networks supposedly evolved more random smaller modules connected by various inter-modular links, rather than stronger larger modules with few interconnections. Also, the evolution of modular structure appeared better consolidated by pairwise (PX and SPX) and to a lesser degree composability (CX) constraint rather than adjacency (PAX and SPAX) restriction. Comparing patterns of modularity of evolving networks to those of randomness (given by RVN_p -value) indicated complementary transitions between the two behaviors over the evolutionary timeline (Figs. 2.4 and 2.5).

In order to dissect the modular behavior of evolving networks, we studied the regression patterns of C against network size N and evolutionary age nd . For typical scale-free models, C declines sharply with increasing N ($C \sim N^{-\text{coefficient}}$), while the coefficients are as high as 0.75 (Albert and Barabási, 2002). Instead, highly modular networks are typically independent of N (Ravasz *et al.*, 2002). In our networks, C regressed by N with very low coefficients (CX, 0.000036; PX, 0.00007; PAX, 0.000035; SPX, 0.00016; SPAX, 0.00016). In contrast, the regression of C with age ($C \sim nd^{-\text{coefficient}}$) produced significantly higher coefficients (CX, 0.39; PX, 0.85; PAX, 0.39; SPX, 0.35; SPAX, 0.41) (Fig. 2.5). The reference power-law (Barabási) networks that were used as control showed a C of zero, as expected (Newman, Strogatz and Watts, 2001). Our data strongly suggests the existence of a highly modular structure that is independent of network growth but is strongly constrained by history, especially when considering the pairwise interactions of the PX network. The rise of the modularity index with emerging power-law degree distribution during certain periods of network evolution indicated a parallel formation of complex hierarchical module clusters with scale-free properties, not distinct from those present in metabolic networks (Ravasz *et al.*, 2002). Our networks of domain organization show a slight lag between an onset of scale-free organization (measured with KS fit and γ indegree statistics) and a delayed emergence of modular behavior (measured with C), occurring during early protein evolution. This was followed by intermittent periods of hierarchical modularity spanning across the middle of the evolutionary timeline. Remarkably, the evolving networks showed a prominent biphasic pattern of hierarchical modularity involving two peaks of modularity (higher statistic C) coinciding with increased power-law behavior (valleys of KS fit and $-\gamma$ curves), at $nd \sim 0.37$ and $nd \sim 0.67$, respectively (Figs. 2.4 and 2.5). The modularity heatmaps and dendrograms of select phases of network evolution confirm these biphasic patterns (Fig. 2.6), which were markedly distinct from the long-

tailed clustering patterns of preferential attachment (Supplementary Fig. S10). As identified earlier (Aziz, Caetano-Anollés and Caetano-Anollés, 2016), the timing of this switch coincides with the early development of genetic code specificity in the emerging ribosomal aaRS catalytic domains, which was facilitated by the OB-fold structure (Caetano-Anollés, Wang and Caetano-Anollés, 2013). These counteracting and delicately balanced trends of modularity and preferential attachment suggest that the emergence of scale-free behavior of the partial bipartite CX network must have impacted the hierarchical modular structure of the modern pairwise networks of domain organization (PX, PAX, SPX, SPAX) (Supplementary Video 3). A detailed account of our testing and verification of this conjecture is explained in Section 7 of Supplementary Text.

2.4 Conclusions

We here present the first tracing of evolutionary ages of protein domain architectures in evolving networks of domain organization. These growing networks revealed major evolutionary pathways of molecular recruitment of domains and functions. Two prominent ancestral waves of structural novelty involved ancient domain innovations and founder ‘*p*-loop’ and ‘winged helix’ domain structures. We found that the evolutionary recruitment of domains and multidomains in proteins is ongoing and highly modular. Remarkably, the networks highlighted the role of cofactor-supporting structures of NRPS pathways, which were backbone to the early evolution of the genetic code. The evolving domain rearrangements featured multitier evolutionary episodes of scale-free network structure, hierarchy and modular behavior. Remarkably, our analyses support biphasic patterns of diversification and module emergence that we have observed earlier (Mittenthal, Caetano-Anollés and Caetano-Anollés, 2012; Aziz, Caetano-Anollés and Caetano-Anollés, 2016). In an initial phase, at the cusp of architectural diversification, the modular components of emerging domain organization associated through weak linkages of recruitment.

The second phase was massive and prolonged, with a multitude of modules appearing after the ‘big bang’ of the protein world, supporting the onset of organismal diversification. Such biphasic patterns are prevalent in biology and impact size, dipeptide makeup, and loop-mediated flexibility of proteins, possibly due to their intrinsic disorder (Caetano-Anollés, Wang and Caetano-Anollés, 2013; Debès *et al.*, 2013). We propose that biphasic patterns in evolving networks are integral to biological history.

2.5 Methods

2.5.1 Experimental design

2.5.1.1 Phylogenomic analysis of the entity set of protein domain architectures

We explore the evolution of networks describing how structural domains combine and split to form single domain and multidomain proteins, i.e. the domain organization of proteins. The definition of protein domain structures followed the FSF level of SCOP version 1.75 (Murzin *et al.*, 1995) (Fig. 2.1). Domain interactions were studied along an evolutionary timeline of structural and architectural innovation directly derived from a phylogenomic tree of architectures reconstructed from an HMM-based census of structural domain organization of 1,730 FSF structures from ~3 million protein sequences encoded in 749 genomes of 52 archaeal, 478 bacterial and 219 eukaryal organisms (dataset A749) (Wang, Kurland and Caetano-Anollés, 2011) (Supplementary Fig. S1). The percentage of proteins with structural assignments was 62.2 ± 0.09 (SD)% (see Table S1 of and discussion in ref. (Wang, Kurland and Caetano-Anollés, 2011)). The tree was generated using maximum parsimony as the optimality criterion in PAUP* following the parsimony ratchet search strategy described in ref. (Wang and Caetano-Anollés, 2009). Data matrix and tree files are provided as Supplementary Data Files 1-3 at <https://github.com/gcalab/SciRep>. The phylogeny represents reconstruction of the “natural

history” of proteins that is supported by a model of protein structural growth(Tal *et al.*, 2016) and is carefully indexed with various evolutionary epochs of the protein world(Wang and Caetano-Anollés, 2009).

2.5.1.2 Calculation of the ages of domain organization

The ages of domains and domain combinations were calculated as node distance (*nd*) values, which were derived directly from the rooted phylogenomic tree of protein domain organization(Wang and Caetano-Anollés, 2009). *nd* values describe relative ages (in a relative 0-1 scale) of first appearances of 6,162 domains and domain combinations (multidomains) defined at SCOP FSF level (the extant ‘entity set’ sampled by our study; Fig. 2) Collectively, ages defined an evolutionary timeline embodying architectural transformations and molecular transitions mediated by fusion and fission processes in the form of 169 unique ‘time events’ (age groups or time slivers) (Supplementary Fig. S2). A Python script was used to count the number of nodes from the root (base) of the tree to each leaf node and present the distance matrix of nodes in a relative zero-to-one scale(Aziz, Caetano-Anollés and Caetano-Anollés, 2016). The script utilized the high imbalance of phylogenomic trees as a fundamental feature to derive the relative ages of domain organization(Wang and Caetano-Anollés, 2009). The tree imbalance resulted from the accumulation of structures and their combinations in proteins and proteomes and not from node density, thus representing a true evolutionary process(Wang *et al.*, 2011).

The timeline was calibrated with a molecular clock of FSF structures ($t = -3.831nd + 3.628$) used to calculate geological age in Gy through calibration points of FSF domains associated with microfossil, fossil and biogeochemical evidence, biomarkers, and first-appearance of clade-specific domains(Wang *et al.*, 2011). The *RSCB – PDB* count was determined by following the

hyperlink associated to the number of entries or structures (which is updated weekly) and selecting “Customizable Table” from the ‘Reports’ menu above the results section. Subsequently, SCOP, CATH, and PFAM ID options were selected as domain information under the ‘Domain Details’ section and domain counts data were exported as a comma separated value (.csv) file report. Supplementary Tables 1-5 provide an exhaustive summary of various connectivity categories of evolving networks based on this ‘entity set’ of domain organization. The extraction pipeline of SPX/SPAX domain units from the original data set can be found in Supplementary Table 6.

2.5.1.3 Indexing domain attributes

Domain ages and assignment of fusional/fissional properties followed ref.(Wang and Caetano-Anollés, 2009). SCOP *concise classification strings (ccs)* of domain descriptions(Murzin *et al.*, 1995) were downloaded from <http://scop.mrc-lmb.cam.ac.uk/scop/parse/index.html> for SCOP version 1.75 as the file `dir_des_scop_txt_1_75.txt`. Available descriptions for 2,223 single domains were obtained from SCOP unique identifiers (sunID). The Gene Ontology (GO) specifications were recorded from the Superfamily Database (SUPFAM) available at <http://supfam.cs.bris.ac.uk/SUPERFAMILY/GO.html>. High-coverage domain-centric GO annotations that were supported only by all UniProts (including multidomain UniProts) were downloaded as the file `Domain2GO_supported_only_by_all.txt`. High-quality truly domain-centric GO annotations that were supported by both single domain UniProts and all UniProts (including multidomain UniProts) were downloaded as the file `Domain2GO_supported_by_both.txt`. We reported only the GO annotations ‘by all’ to capture higher coverage. Also, the GO terms were reported only for the 2,223 single domains with descriptions available. Specialized GO annotations from two levels of hierarchy downstream were

taken from files Domain2GO-Hie-Dist1.csv and Domain2GO-Hie-Dist2.csv. Structural domains functional ontology (SDFO) that mapped information from a theoretic analysis of Domain2GO annotation profiles were reported from the file SDFO.txt.

2.5.1.4 Network construction, visualization and analysis

Mathematical definitions for construction of networks can be found in Supplementary Materials and Methods. The social network analysis tool Pajek(Mrvar and Batagelj, 2016) and the statistical test bench R's *igraph* package(Csardi and Nepusz, 2006) were used to visualize and analyze the networks, respectively. The collective impact of events was made explicit by Pajek's Visualization of Similarity (VOS) clustering method(Van Eck and Waltman, 2007; Waltman, van Eck and Noyons, 2010). VOS helped reveal communities and design layouts of networks with nodes separated into network modules, where high *modularity indices* ranged from 94-95%. Number of clusters varied over networks (CX, 691; PX, 3,886; PAX, 4,126; SPX, 607; SPAX, 620). Network clusters were visually compacted to hubs and their cohesive neighbors with the energy-optimizing Kamada-Kawai 'separate components' algorithm(Kamada and Kawai, 1989). Pajek allowed to proportionally reduce the size of highly connected nodes by some scaling factor for optimally uncluttered visualization. Waterfall and radial network layouts were designed with node-size scaled down by factors of 0.1 and 0.25, respectively. R packages equipped with specialized code constructs to draw graphs and derive statistics were used to analyze network properties(Ihaka and Gentleman, 1996; R Core Team, 2014). We also used Pre-Hypertext Processing language (PHP) to write custom scripts that generated radial visualizations of the networks and helped conduct housekeeping data management(PHP-Group and others, 2012). The PHP scripts were executed in the command line. Results of these scripts were input into Pajek's and R's analytical procedures.

We used the open-source software ImageMagick (www.imagemagick.org) for batch conversion, captioning, and appending of network images (to represent legends and scales). A detailed description of partition and data files, list of network data analysis functions, charting and graphing procedures, methods to generate power law statistics, modularity indices and randomness checks, and the method pipeline used to achieve waterfall diagrams can be found in Supplementary Materials and Methods.

2.5.2 Statistical analysis

2.5.2.1 Scale-free network behavior

Linear regression models of $P(k)$ given k (i.e. the probability of having k -neighbors) were used to derive the γ coefficient of the power law distribution and the determination coefficient, R^2 . The value of γ represents an absolute slope of the log linear model of $P(k)$ vs. k . The slope is usually ≤ 0 . $\gamma \gg 1$ indicates strong tendency towards preferential attachment. R^2 indicates the percentage of data that fits the linear model. High values of both γ and R^2 suggest strong scale-free behavior. Additional power law statistics were calculated as: (i) the exponent of the fitted power law distribution, α , with an assumption that $P(X=x)$ is proportional to $x^{-\alpha}$; (ii) KS fit statistic to compare the input degree distribution with that of fitted power-law; and (iii) the KS p -value of a statistical test, with the null hypothesis that data is being drawn from a power law distribution (Newman, 2005; Clauset, Shalizi and Newman, 2009). $\alpha \gg 1$, $0 < \text{KS fit scores} \ll 1$, and KS p -values ≥ 0.05 suggest that degree data was derived from a fitted power law distribution. Maximum log likelihood of the fitted scale-free parameters was also determined. Control networks were included for reference that were generated with ‘Barabási’ methods (Barabási and Albert, 1999) of the *igraph*

package from R(Csardi and Nepusz, 2006). These controls simulated basic and extended age-dependent power law graph models given varying sizes of the evolving networks.

2.5.2.2 Network modularity

We investigated modularity using six indices: (i) The VOS Quality index (VQ) was determined using the Pajek VOS algorithm by considering the number or weights of the links (arcs) between the nodes as similarities. Clusters or communities that were deemed ‘similar’ were iteratively drawn closer to each other until a final layout was achieved with least crossings and closest clusters. The quality index VQ was thus calculated for this final layout as $\sum_{i=1}^c \sum_{j=i+1}^c (e_{ij} - a_i^2)$, where c is the number of communities; e_{ij} is the fraction of edges with one node v in the community i (c_i) and the other node w in the community j (c_j), defined as $\sum_{vw} (A_{vw}/2m)$ where $v \in c_i$, $w \in c_j$, m is the sum of weights in the graph and A_{vw} is the weighted value or 0, indicating presence or absence of edge between nodes v and w in the adjacency matrix A of the network; and a_i is the fraction of weighted k neighbors attached to the nodes in community i , i.e. $k_i/2m$ (Van Eck and Waltman, 2007; Waltman, van Eck and Noyons, 2010). (ii) The Clustering Ratio (C-ratio) is the ratio of the number of network clusters to the count of the connected nodes in the network. (iii) The average Clustering Coefficient (C) is defined as the ratio of the triangles impingent on a node to the connected triples, determined as a global average over all nodes in a simplified (undirected/unweighted) network (Wasserman and Faust, 1994; Ravasz *et al.*, 2002; Barrat *et al.*, 2004). C is not meaningful for strictly bipartite or scale-free graphs (Newman, Strogatz and Watts, 2001). We also report coefficients of linear regression of C over the age and size of the networks of domain organization. (iv) The Fast-Greedy Community (FGC) agglomerative hierarchical algorithm detects community structure for networks with m edges, n nodes, and a depth d of the

dendrogram describing the community structure, given an optimized linear running time of $O(m \times d \times \log n) \sim O(n \times \log^2 n)$ (Clauset, Newman and Moore, 2004). The *Newman-Girvan* algorithm index (*NG*) was computed with two different input partitions, the first (v) defined by age (NG_{age}) and the second (vi) defined by VOS clustering (NG_{vos}). *NG* calculates the modularity of a network given a predefined division or partition to measure the influence of the partition in separating the different node types. This indicates either assortative (positive) or disassortative (negative) mixing across modules (Newman and Girvan, 2004). The *NG* algorithm computes an index as $1/(2m) \sum_{ij} (A_{ij} - 1/(2m) k_i k_j \times \Delta(c_i, c_j))$, where m is the sum total of weights in the graph and A_{ij} are weighted entries in the adjacency matrix of the network; k_i | k_j and c_i | c_j are the weighted degrees and the components (numeric partitions) of the nodes i and j , respectively; finally, $\Delta(x, y)$ equals 1 if $x=y$ and 0 otherwise (Newman and Girvan, 2004). The *VQ*, *C-ratio*, *C* and *FGC* indices each range from 0 to 1, while the *NG* indices range from -1 to 1. In all cases, higher values represent strong modularity of the network at an event of evolutionary history. Heatmaps of modularity were constructed using log10-scaled modularity matrices, with each map element given as $(A_{ij} - k_i k_j / (2m)) / M_{nd}$, where A_{ij} , k_i , k_j and m were the same as defined for *NG* (Newman and Girvan, 2004), while M_{nd} was the network's modularity index at event nd . Cladistic representations of modularity were visualized with dendrograms whose metrics were calculated from squared Euclidean distance matrices, which indicate dissimilarities between cluster means (Borg and Groenen, 2003). The dissimilarity or distance matrices were clustered hierarchically using the Ward's minimum variance method that seeks compact and spherical clusters (Murtagh and Legendre, 2014).

2.5.2.3 Quantifying randomness in networks

The Bartels rank test of randomness, which primarily offers a rank version of von Neumann's Ratio Test for Randomness(Bartels, 1982), was used to measure random network behavior. The resultant test statistic RVN is defined as $\sum_{i=1}^{n-1} (R_i - R_{i+1})^2 / \sum_{i=1}^n (R_i - (n+1)/2)^2$, where $R_i = \text{rank}(X_i)$ with $i=1 \dots n$, $(RVN-2)/\sigma$ is the asymptotically standard normal, and $\sigma^2 = [4(n-2)(5n^2-2n-9)]/[5n(n+1)(n-1)^2]$. The null hypothesis of this method was randomness, which was tested against the alternate hypothesis of non-randomness, given a trend of RVN values. A p-value is computed from a two-sided beta distribution approximation test. Random graph controls were created by following the Erdős–Rényi graph model(Erdős and Rényi, 1959, 1960).

2.6 Figures

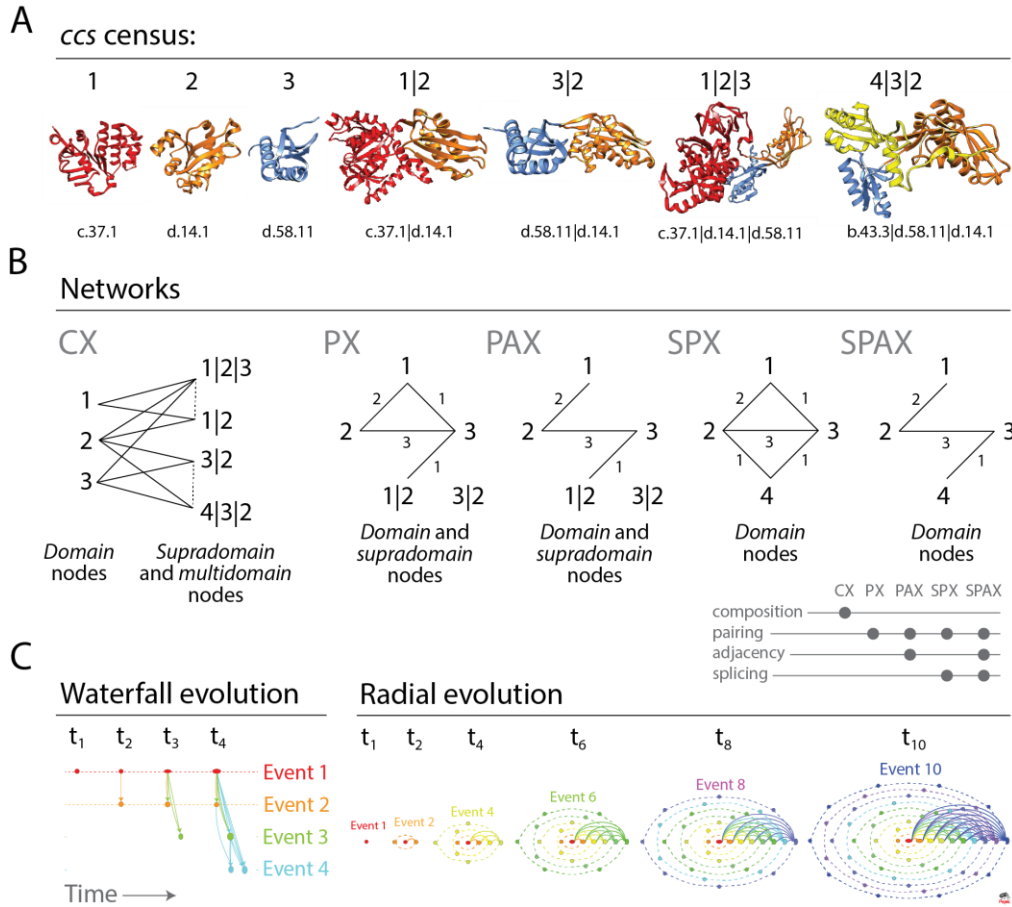


Figure 2.1. Networks of protein domain organization. (A) The genomic census of structural domains and their combinations defines SCOP *concise classification string* (*ccs*) descriptors of domains, supradomains and multidomains that are building blocks of networks. We illustrate the census with a sample from the entire entity set, comprising of 3 domains (1, 2 and 3), 2 supradomains (1|2 and 3|2) and 2 multidomains (1|2|3 and 4|3|2) that are common in dehydratase enzymes and elongation factors. *ccs* identifiers of structural domain constituents defined at fold superfamily (FSF) level are listed below the atomic models visualized in ribbon format with Chimera. (B) Five operative criteria for network generation capture the interactions among protein architecture nodes as networks grow in evolution. CX is a partial bipartite network (projection-decomposable) that connects domain nodes to supradomain and multidomain nodes (which can connect to each other; hatched links) when present in multidomain proteins. PX connects domain and supradomain nodes when multidomain proteins are ‘decomposed’ into pairs of architectures, regardless of topological constraints. PAX borrows the PX criterion but respects topological constraints. SPX connects domain nodes spliced from architectures when domain pairs are present in proteins. SPAX connects domain nodes when adjacent domain pairs are present in proteins. (C) Chronological development of evolving networks. In ‘waterfall evolution’ layout, time progresses from left to right as ‘discrete events’ of network evolution progressively unfold the appearance of nodes and links (time-directed arrows known as arcs) from top to bottom, colored according to their age. Arc multiplicities describe link cardinality. Source-sink recruitments of architectures are visualized by horizontal and vertical elongations of node symbols, which describe their outdegree and indegree, respectively. As networks grow, the symbols of older nodes widen by outdegree accumulation, while those of younger nodes grow tall by indegree accumulation. In ‘radial evolution’ layout, the time-variant network grows by accumulating nodes in

Figure 2.1 (cont.)

concentric rings (orbitals), each reflecting a time event. We illustrate radial evolution with 6 snapshots of a network growing to a size of 55 nodes as it unfolds from time t_1 to t_{10} . Nodes (n) in orbitals (r) grow at $r+1$ rate and only one node per orbital connects to single nodes in each of the other orbitals. Thus, outward links (o) of an orbital are $o=t-r-1$, where t is the current time. Inward links (i) of an orbital are $i=t-o-1=r$. Finally, total links of a network at any time are $t(t-1)/2$. The width and height of symbols represent the outdegree and indegree of nodes, respectively. Symbol sizes are shifted by 10 for a better visualization of nodes.

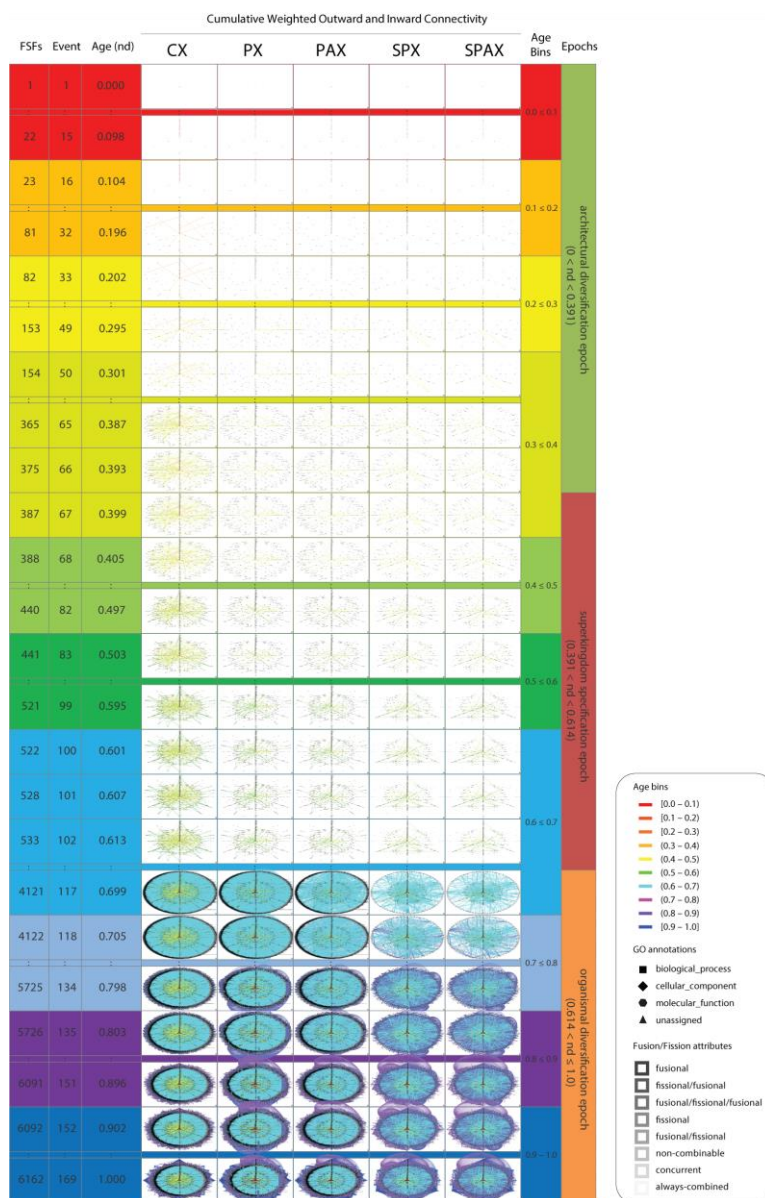


Figure 2.2. Evolving networks in radial evolution layout. Snapshots of network growth describe the evolution of 6,162 domain, supradomain and multidomain architectures or 1,643 domains spliced from them. They represent 24 out of 169 time events of the evolutionary timeline, which are indexed with evolutionary age (nd, ranging from 0.0 to 1.0), age bin (one of 10), and one of the 3 epochs of protein evolution (Wang *et al.*, 2007). Age bins were custom RGB color-coded to highlight the flow of time, from top to bottom. The evolving CX, PX, PAX, SPX and SPAX networks reveal the gradual evolutionary accumulation of nodes and links. The sizes of the horizontal and vertical axes of the node symbols depict outward and inward weighted connectivity, respectively, with all weighted degree vectors shifted by 10 for visualization and inclusion of 0-degree nodes. The curved arcs describe recurring interactions between architectures that are accumulating along the successive events of the timeline. Arcs symbolize the flow of time from ancient to recent architectures and are color-coded according to the age of the more recent of the component nodes involved; arcs between contemporary nodes are excluded. Since, in pairwise networks the age of the most recent parent node could be assigned to the arc, the connectivity-defining pairing events are absent in the first (red) and the first and second (red, orange) bins of the PX and SPX and the PAX and SPAX networks, respectively. The angles of multiple arcs emerging from nodes are incremented by 2 to avoid overlap. Node RGB colors represent age. Grey-scale color of node borders depict fusional/fissional properties. Node shapes describe GO categories. High-resolution figures and Pajek network files are available at <https://github.com/gcalab/SciRep>.

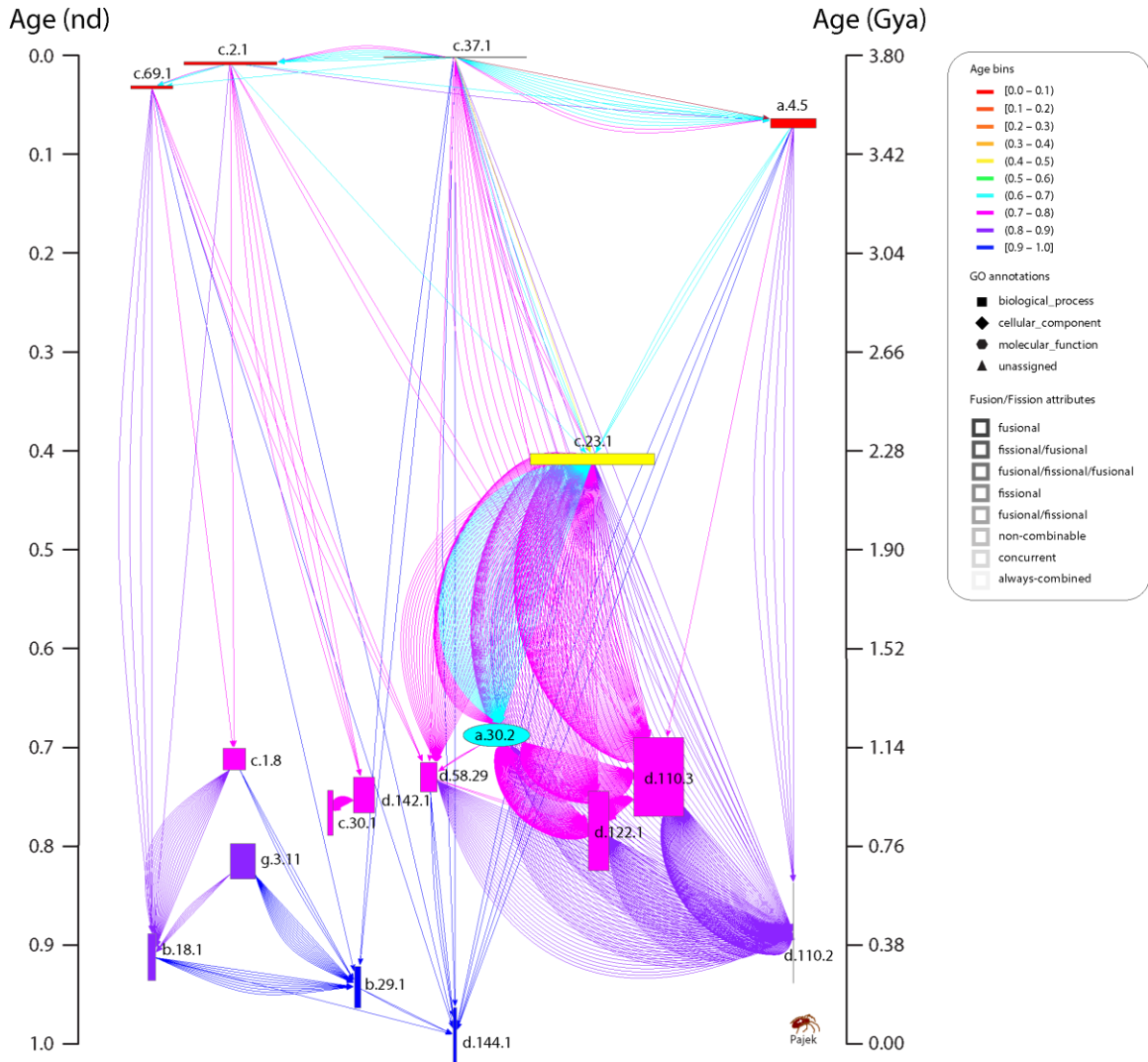


Figure 2.3. An extant SPX network in waterfall layout describing the evolution of spliced domains with the largest (100th percentile) network connectivity. The SPX network of 1,643 spliced domains was reduced with the restrictive criterion of excluding nodes with combined outdegrees and indegrees $\geq 99\%$ of those of the rest of the nodes. The set of arcs (arched arrows symbolizing flow of time) was also reduced to pairing events between domains in the 100th percentile connectivity and excluded those between contemporary nodes. Nodes are arranged top-down and colored according to age (*nd*) on a relative 0-to-1 scale that describes evolutionary time events. Ages are also time-calibrated with a molecular clock of FSF domains, which uses fossils and microfossils, geochemical, biochemical, and biomarker data (Wang *et al.*, 2011). FSF origin is given in billion years ago (Gya). Nodes were labeled with SCOP ccs domain descriptors. To showcase source-and-sink relationships, node symbol sizes were scaled proportional to the weighted outdegree and indegree along the horizontal and vertical axes, respectively. Weighted degrees were scaled as $\times 2+2$ to include 0-degree nodes for better visualization. The modular spread of nodes was based on VOS clustering (see methods). Arcs are color-coded according to the age of the more recent of the component nodes involved; no arcs were present in the ancient-most age bin (red) of the timeline. Angles of multiple arcs emerging from nodes are incremented by 2 to avoid overlap.

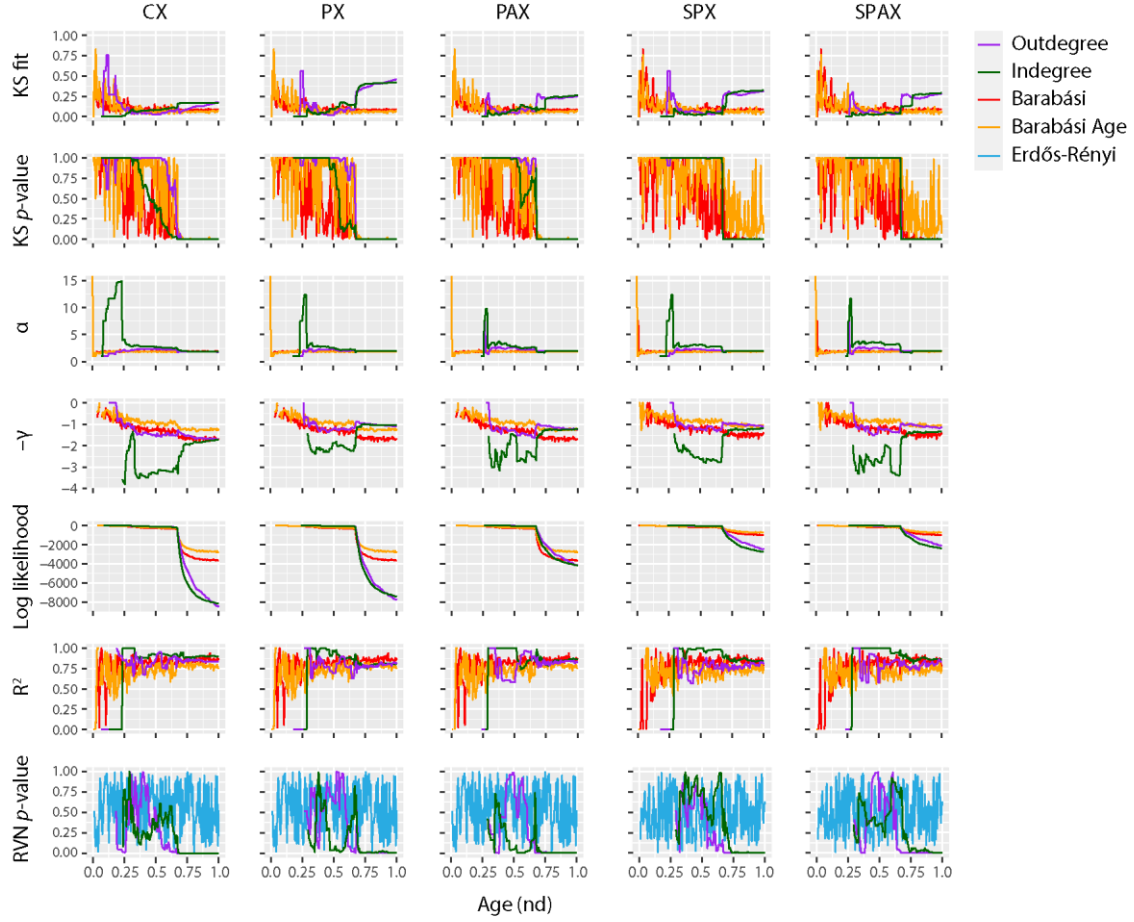


Figure 2.4. Statistical descriptors of power law and random behavior. Six indicators of preferential attachment were studied along the evolutionary timeline to explore processes of network growth, with network age (nd) indicated on a relative 0-to-1 scale. Outdegree and indegree connections were cumulative and weighted in evolving networks. Barabási (red) and Barabási-Age (orange) networks were included as control sets. The Barabási model specifies the probability of preference of an old node as $P_i \sim k_i^\alpha$ while the Barabási-Age model grants heavier power law properties to older nodes (exhibiting smaller nd) with $P_i \sim (k_i^\alpha)(l_i^\beta)$, where k_i is the indegree of node i of the current event, α is the preferential attachment exponent ($\alpha = 1$ for linear preferential attachment), l_i is the age of node i , i.e. the number of events elapsed since the node was added, with maximum number measured by the ‘aging.bin’ parameter, and β is the aging exponent ($\beta = 1$ for linear increases in probability of preference of an older node with high l_i). Power law indices include: (i) the KS fit statistic that compares the input degree data distribution with the fitted power law distribution (smaller scores denote better fit); (ii) the KS p-value, which rejects the null hypothesis that degree data was drawn from the fitted power-law distribution when less than $\alpha=0.05$; (iii) the exponent of the fitted power-law distribution (α); (iv) the slope of power-law linear regression model (γ); (v) the log-likelihood of the fitted parameters; and (vi) the coefficient of determination (R^2) that measures the percentage of degree data that fits the linear model. The randomness of the evolving networks was quantified by the p-value of an approximated beta distribution from the rank version of von Neumann's Ratio Test for Randomness (Bartels, 1982) ($RVN_{p\text{-value}}$). The alternate hypothesis was non-randomness. Comparative graphs of strictly random Erdős–Rényi control networks of corresponding sizes at the given time-events were also plotted. Lower KS fit, higher KS p-value, higher α , lower $-\gamma$ and near-zero likelihood, given lower $RVN_{p\text{-value}}$, support power law behavior.

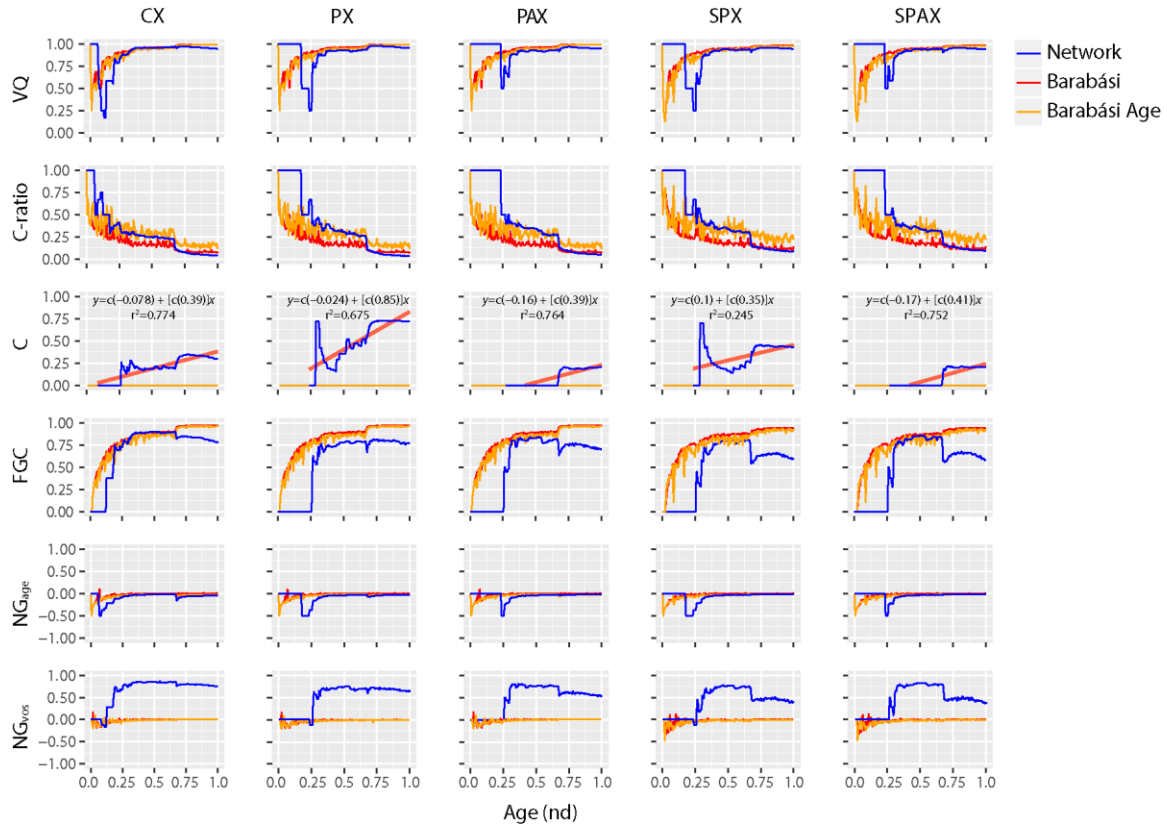


Figure 2.5. Network modularity. Six indicators of modularity were studied along the evolutionary timeline to explore the evolution of network structure, with network age (nd) indicated on a relative 0-to-1 scale. Modularity indices include the VOS Quality (VQ) index, the Clustering ratio (C-ratio), the average Clustering Coefficient (C), the Fast-Greedy Community (FGC) index, and the Newman-Girvan index defined by age (NG_{age}) or VOS clustering (NG_{vos}). Modularity calculations required cumulative, undirected, and weighted connectivity input. The Barabási (red) and Barabási-Age (orange) models (see caption of Figure 2.4) were included as control sets. The regressions of C with age (nd) are shown as linear models (red lines) for each network together with supporting determination coefficients (R^2).

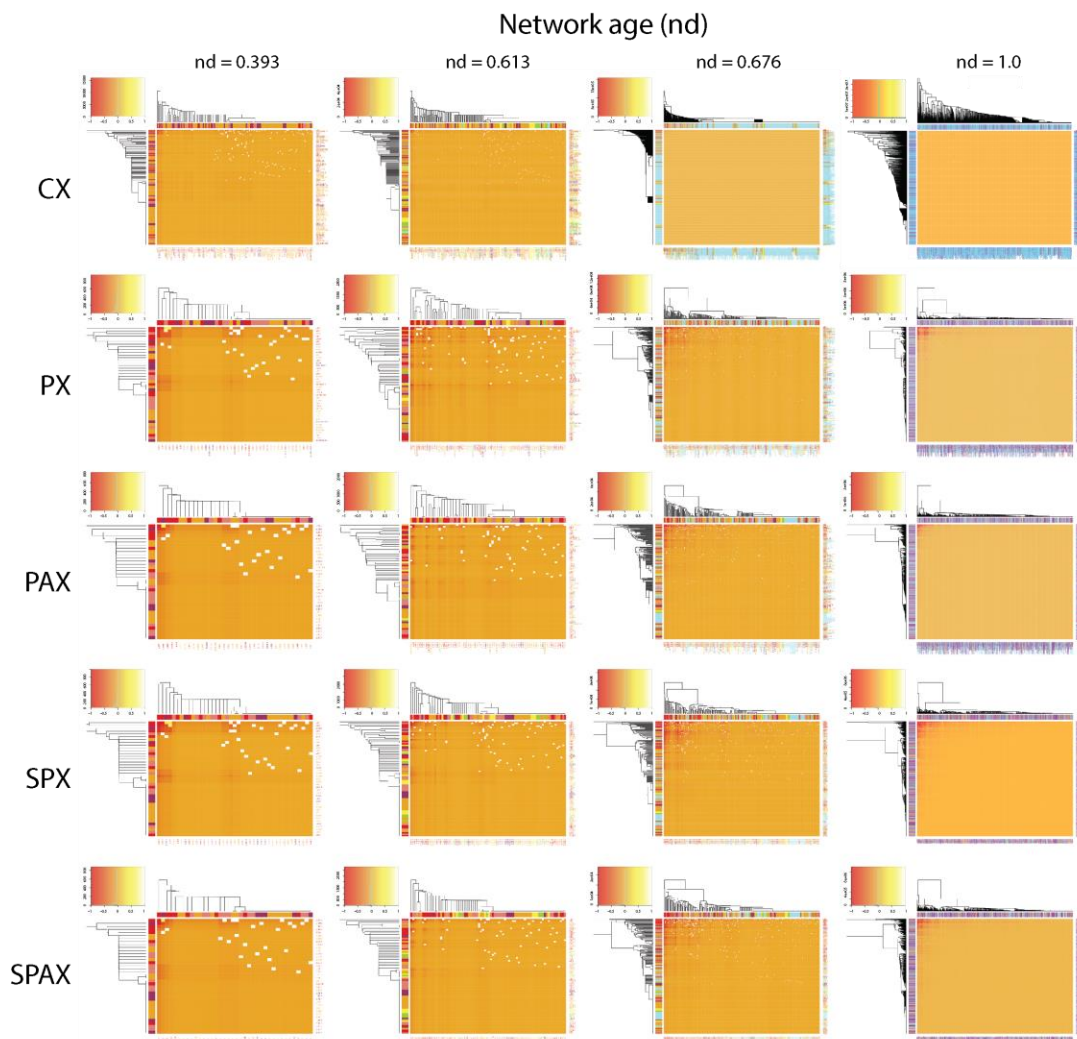


Figure 2.6. Evolution of modularity and hierarchical organization of networks over select events of the evolutionary timeline. NG_{age} pairwise modularity values (Newman and Girvan, 2004), scaled by log₁₀ of network-wide absolute modularity values, were used as input for the calculation of Euclidean distance matrices (Borg and Groenen, 2003), which were visualized as heatmaps. Heatmap tiles represent modular strength between any two architectures relative to the respective strength of their linkages to other architectures of the network. The embedded dendrograms that define the order of rows and columns of the heatmaps were generated by hierarchical clustering of the distance matrices with the Ward's minimum variance method (Murtagh and Legendre, 2014). The height of dendrograms represents dissimilarity between clusters while the clades show grouping rearrangements of architectures. The top-left insets depict frequency histograms of the heatmap modularity values scaled from -1 to 1 (i.e. disassortative to assortative). The four panels describe growth of each evolving network (left-to-right). Network age corresponds to the middle approximate boundaries of the three evolutionary epochs of the protein world (Supplementary Fig. S2), i.e., end of 'architectural diversification' (nd = 0.393), end of 'superkingdom specification' (nd = 0.613), onset of the 'big bang' of domain organization at the start of 'organismal diversification' (nd = 0.676); and the present (nd = 1). Nodes were age-sorted ascendingly within clusters and labelled using standard SCOP nomenclature (Murzin *et al.*, 1995). In the case of SPX and SPAX, nodes correspond to 1,643 domains mapped to the entity set of 6,162 architectures. The color-coding of bands and labels identifies the age of architectures (Supplementary Fig. S2). The relatively 'flatter' heatmap and 'skewed' dendrogram patterns of CX (typically at nd = 0.667 and nd = 1.000) are an artifact of unweighted distance matrices of CX, which contrast with the weighted ones of pairwise criterion-based networks. The most prominent clades correspond to the modules of the most ancient domain

Figure 2.6 (cont.)

structures harboring the two major waves of architectural innovation. We also generated heatmaps of power-law control networks of corresponding sizes at the given time-events (Supplementary Fig. S10). When compared to the pairwise networks, the combined heatmap and dendrogram patterns of CX suggest a hidden switch from scale-freeness to modular behavior, eventually giving rise to hierarchical modularity with visible emergence of modules within modules.

2.7 Tables

Table 2.1: Domains and domain combinations scoring $\geq 99.9^{\text{th}}$ percentiles of 249.916, [64] and {23}, based on combined outdegrees of the five networks at time points 1.0, [0.676] and {0.671}, respectively. The square and curly brackets denote values from the events after and before the big bang, respectively.

Age Rank	Label	Node Age	Network(s)	Out Degree	Fusional / Fissional	Description	GO Name
388	c.23.1	0.4046243	PX, PAX, SPX	1013, 390, 330	fissional/fusional	CheY-like	regulation of multicellular organismal development
1	c.37.1	0.0000000	PX, SPX, PAX, SPAX, CX	607, 380, 376, 314, 271	fusional	P-loop containing nucleoside triphosphate hydrolases	positive regulation of reproductive process
			[CX, PX, SPX, PAX, SPAX]	[109, 97, 74, 64, 64]			
			{CX, PX, SPX, PAX, SPAX}	{34, 32, 26, 23, 23}			
2446	a.30.2	0.6820809	PX	578	fissional/fusional	Homodimeric domain of signal transducing histidine =	alkene binding
48	e.23.1	0.1445087	PX	452	fusional	Acetyl-CoA synthetase-like	regulation of primary metabolic process
			[PX]	[75]			
2	c.2.1	0.0057803	PX	427	fusional	NAD(P)-binding Rossmann-fold domains	pyridine-containing compound metabolic process
			[PX, CX, SPX]	[117, 80, 66]			
1518	d.110.3 a.30.2	0.6763006	PX	423	fissional/fusional	#N/A	#N/A
283	a.28.1	0.3526012	PX	416	fusional	ACP-like	cell periphery
2543	d.110.3&	0.6820809	PX	369	fissional/fusional	#N/A	#N/A
858	d.110.2 d.110.3	0.6763006	PX	357	fissional/fusional	#N/A	#N/A
187	c.30.1 d.142.1	0.3179191	PX	352	fissional/fusional	#N/A	#N/A
8	c.69.1	0.0346821	PX	327	fusional	alpha/beta-Hydrolases	regulation of multicellular organismal development
4777	d.110.3	0.7225434	PX	325	fissional/fusional	PYP-like sensor domain (PAS domain)	regulation of cellular macromolecule biosynthetic process
17	c.23.16	0.0809249	PX	315	fusional	Class I glutamine amidotransferase-like	positive regulation of oxidative phosphorylation uncoupler activity
4465	d.122.1 c.23.1	0.7109827	PX	291	fusional/fissional/fusional	#N/A	#N/A
1599	d.110.3 d.110.2	0.6763006	PX	262	fissional/fusional	#N/A	#N/A
443	c.1.33	0.5028902	PX	253	fusional	EAL domain-like	cyclic-guanylate-specific phosphodiesterase activity

Table 2.2: Domains and domain combinations scoring $\geq 99.9^{\text{th}}$ percentile of 247,977, [21] and {5}, based on combined indegrees of the five networks at time points 1.0, [0.676] and {0.671}, respectively. The square and curly brackets denote values from the events after and before the big bang, respectively.

Age Rank	Label	Node Age	Network(s)	In Degree	Fusional / Fissional	Description	GO Name
6044	d.110.2	0.8728324	PX, PAX, SPX	766, 295, 267	fissional	GAF domain-like	purine-containing compound catabolic process
4777	d.110.3	0.7225434	PX	735	fissional/fusional	PYP-like sensor domain (PAS domain)	regulation of cellular macromolecule biosynthetic process
5529	d.122.1	0.7745665	PX	701	fissional/fusional	ATPase domain of HSP90 chaperone/DNA topoisomerase =	nucleic acid metabolic process
5038	a.30.2 d.122.1	0.7341040	PX	550	fusional/fissional/fusional	#N/A	#N/A
5101	c.43.1	0.7398844	PX	445	fissional/fusional	CoA-dependent acyltransferases	monocarboxylic acid catabolic process
5664	d.110.3 a.30.2 d.122.1	0.7919075	PX	439	fusional/fissional	#N/A	#N/A
6150	c.43.1&	0.9768786	PX	432	fusional/fissional	#N/A	#N/A
5304	c.30.1	0.7572255	PX	375	fissional/fusional	PreATP-grasp domain	pyrimidine-containing compound biosynthetic process
6148	b.1.1	0.9768786	PX	370	fissional	Immunoglobulin	regulation of mesoderm development
4848	e.23.1 a.28.1	0.7225434	PX	367	fusional/fissional	#N/A	#N/A
5095	d.142.1	0.7398844	PX	359	fissional/fusional	Glutathione synthetase ATP-binding domain-like	pyrimidine-containing compound biosynthetic process
5731	g.3.11	0.8034682	PX	317	fissional/fusional	EGF/Laminin	positive regulation of receptor activity
4118	c.43.1& e.23.1 a.28.1	0.6994219	PX	287	fusional/fissional/fusional	#N/A	#N/A
4758	d.58.29	0.7225434	PX	272	fissional/fusional	Nucleotide cyclase	regulation of primary metabolic process
5521	d.110.3 d.58.29	0.7745665	PX	266	fusional/fissional	#N/A	#N/A
4855	c.43.1& e.23.1 a.28.1 c.43.1&	0.7225434	PX	265	fusional/fissional	#N/A	#N/A
4763	a.30.2 d.122.1 c.23.1	0.7225434	PX	263	fusional/fissional/fusional	#N/A	#N/A
5768	b.23.1	0.8092486	PX	261	fissional/fusional	Spermadhesin, CUB domain	regulation of anatomical structure size
5759	b.1.2	0.8092486	PX	259	fissional/fusional	Fibronectin type III	regulation of CD4-positive, alpha-beta T cell activation
2886	c.43.1 e.23.1 a.28.1	0.6878613	PX	258	fusional/fissional/fusional	#N/A	#N/A
[1620]	c.43.1& e.23.1	0.6763006	PX	28	fusional/fissional/fusional	#N/A	#N/A
[1223]	d.142.1 c.24.1	0.6763006	PX	25	fusional/fissional/fusional	#N/A	#N/A
[1311]	a.28.1 c.43.1&	0.6763006	PX	25	fusional/fissional/fusional	#N/A	#N/A
[1032]	e.23.1 a.28.1 c.43.1& e.23.1	0.6763006	PX	23	fusional/fissional/fusional	#N/A	#N/A
[283]	a.28.1	0.3526012	PX, SPX, PAX, SPAX	22, 22, 21, 21	fusional	ACP-like	cell periphery

Table 2.2 (cont.)

Age Rank	Label	Node Age	Network(s)	In Degree	Fusional / Fissional	Description	GO Name
[1556]	d.142.1 a.92.1 c.30.1 d.142.1 c.24.1	0.6763006	PX	22	fusional/fissional/fusional	#N/A	#N/A
[1085]	e.23.1 a.28.1 c.43.1&e.23.1 a.28.1	0.6763006	PX	21	fusional/fissional/fusional	#N/A	#N/A
{672}	a.4.1	0.6647399	PX, CX	8, 7	fissional/fusional	Homeodomain-like	regulation of epithelial cell differentiation involved in kidney development
{324}	c.73.1	0.3641618	PX	6	fissional/fusional	Carbamate kinase-like	heterocycle metabolic process
{460}	c.73.1 d.58.18&c.2.1 d.81.1	0.5202312	CX	5	fusional/fissional	#N/A	#N/A
{734}	b.113.1 a.156.1 g.39.1 c.37.1	0.6705202	CX	5	fusional/fissional	#N/A	#N/A
{13}	c.2.1 a.100.1	0.0693642	PX	5	fusional/fissional/fusional	#N/A	#N/A
{58}	d.14.1	0.1791908	PX, SPX	5, 5	fusional	Ribosomal protein S5 domain 2-like	nucleic acid phosphodiester bond hydrolysis
{270}	g.39.1	0.3468208	PX	5	fissional/fusional	Glucocorticoid receptor-like (DNA-binding domain)	fibroblast growth factor receptor signaling pathway involved in ureteric bud formation
{388}	c.23.1	0.4046243	PX	5	fissional/fusional	CheY-like	regulation of multicellular organismal development

2.8 References

1. Chothia, C. & Gough, J. Genomic and structural aspects of protein evolution. *Biochem. J* **419**, 15–28 (2009).
2. Chothia, C., Gough, J., Vogel, C. & Teichmann, S. A. Evolution of the protein repertoire. *Science* (80-.). **300**, 1701 (2003).
3. Wetlaufer, D. B. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. U. S. A.* **70**, 697–701 (1973).
4. Janin, J. & Wodak, S. J. Structural domains in proteins and their role in the dynamics of protein function. *Progress in Biophysics and Molecular Biology* vol. 42 21–78 (1983).
5. Han, J. H., Batey, S., Nickson, A. A., Teichmann, S. A. & Clarke, J. The folding and evolution of multidomain proteins. *Nature Reviews Molecular Cell Biology* vol. 8 319–330 (2007).
6. Apic, G., Gough, J. & Teichmann, S. A. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* **310**, 311–325 (2001).
7. Wang, M. & Caetano-Anollés, G. The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. *Structure* **17**, 66–78 (2009).

8. Caetano-Anollés, G., Wang, M., Caetano-Anollés, D. & Mittenthal, J. E. The origin, evolution and structure of the protein world. *Biochemical Journal* vol. 417 621–637 (2009).
9. Murzin, A. G., Brenner, S. E., Hubbard, T., Chothia, C. & others. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
10. Chandonia, J. M., Fox, N. K. & Brenner, S. E. SCOPe: Manual Curation and Artifact Removal in the Structural Classification of Proteins – extended Database. *J. Mol. Biol.* **429**, 348–355 (2017).
11. Dawson, N. L. *et al.* CATH: An expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.* **45**, D289–D295 (2017).
12. Marchler-Bauer, A. *et al.* CDD: NCBI’s conserved domain database. *Nucleic Acids Res.* **43**, D222–D226 (2015).
13. Bru, C. *et al.* The ProDom database of protein domain families: More emphasis on 3D. *Nucleic Acids Res.* **33**, (2005).
14. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36**, (2008).
15. Finn, R. . *et al.* InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Res* **45**, D190–D199 (2017).
16. Wilson, D. *et al.* SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.* **37**, D380–D386 (2009).
17. Harrison, A., Pearl, F., Mott, R., Thornton, J. & Orengo, C. Quantifying the similarities within fold space. *J. Mol. Biol.* **323**, 909–926 (2002).
18. Berezovsky, I. N., Guarnera, E. & Zheng, Z. Basic units of protein structure, folding, and function. *Prog. Biophys. Mol. Biol.* **128**, 85–99 (2017).
19. Gerstein, M. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold. Des.* **3**, 497–512 (1998).
20. Wang, M., Kurland, C. G. & Caetano-Anollés, G. Reductive evolution of proteomes and protein structures. *Proc. Natl. Acad. Sci.* **108**, 11954–11958 (2011).
21. Illergård, K., Ardell, D. H. & Elofsson, A. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins Struct. Funct. Bioinforma.* **77**, 499–508 (2009).
22. Bashton, M. & Chothia, C. The geometry of domain combination in proteins1. *J. Mol. Biol.* **315**, 927–939 (2002).
23. Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826 (1986).

24. Vogel, C., Berzuini, C., Bashton, M., Gough, J. & Teichmann, S. A. Supra-domains: evolutionary units larger than single protein domains. *J. Mol. Biol.* **336**, 809–823 (2004).
25. Wuchty, S. Scale-free behavior in protein domain networks. *Mol. Biol. Evol.* **18**, 1694–1702 (2001).
26. Apic, G., Huber, W. & Teichmann, S. A. Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. *J. Struct. Funct. Genomics* **4**, 67–78 (2003).
27. Tordai, H., Nagy, A., Farkas, K., Bányai, L. & Patthy, L. Modules, multidomain proteins and organismic complexity. in *FEBS Journal* vol. 272 5064–5078 (2005).
28. Weiner, J., Moore, A. D. & Bornberg-Bauer, E. Just how versatile are domains? *BMC Evol. Biol.* **8**, (2008).
29. Shahzad, K., Mittenthal, J. E. & Caetano-Anollés, G. The organization of domains in proteins obeys Menzerath-Altmann’s law of language. *BMC Syst. Biol.* **9**, (2015).
30. Vogel, C., Teichmann, S. A. & Pereira-Leal, J. The relationship between domain duplication and recombination. *J. Mol. Biol.* **346**, 355–365 (2005).
31. Basu, M. K., Carmel, L., Rogozin, I. B. & Koonin, E. V. Evolution of protein domain promiscuity in eukaryotes. *Genome Res.* **18**, 449–461 (2008).
32. Xie, X., Jin, J. & Mao, Y. Evolutionary versatility of eukaryotic protein domains revealed by their bigram networks. *BMC Evol. Biol.* **11**, (2011).
33. Taylor, W. R. Evolutionary transitions in protein fold space. *Current Opinion in Structural Biology* vol. 17 354–361 (2007).
34. Alva, V., Remmert, M., Biegert, A., Lupas, A. N. & Söding, J. A galaxy of folds. *Protein Sci.* **19**, 124–130 (2010).
35. Ferrada, E. & Wagner, A. Evolutionary innovations and the organization of protein functions in genotype space. *PLoS One* **5**, (2010).
36. Taylor, W. R. A ‘periodic table’ for protein structures. *Nature* **416**, 657–660 (2002).
37. Forslund, K. & Sonnhammer, E. L. L. Evolution of protein domain architectures. in *Evolutionary Genomics* 187–216 (Springer, 2012).
38. Gerstein, M. Patterns of protein-fold usage in eight microbial genomes: A comprehensive structural census. *Proteins Struct. Funct. Genet.* **33**, 518–534 (1998).
39. Abeln, S. & Deane, C. M. Fold usage on genomes and protein fold evolution. *Proteins: Structure, Function and Genetics* vol. 60 690–700 (2005).
40. Edwards, H., Abeln, S. & Deane, C. M. Exploring Fold Space Preferences of New-born and Ancient Protein Superfamilies. *PLoS Comput. Biol.* **9**, e1003325 (2013).

41. Caetano-Anollés, G. & Caetano-Anollés, D. An evolutionarily structured universe of protein architecture. *Genome Res.* **13**, 1563–1571 (2003).
42. Wang, M., Yafremava, L. S., Caetano-Anollés, D., Mittenthal, J. E. & Caetano-Anollés, G. Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. *Genome Res.* **17**, 1572–1585 (2007).
43. Wang, M. & Caetano-Anollés, G. Global phylogeny determined by the combination of protein domains in proteomes. *Mol. Biol. Evol.* **23**, 2444–2454 (2006).
44. Nath, N., Mitchell, J. B. O. & Caetano-Anollés, G. The natural history of biocatalytic mechanisms. *PLoS Comput Biol* **10**, e1003642 (2014).
45. Debès, C., Wang, M., Caetano-Anollés, G. & Gräter, F. Evolutionary optimization of protein folding. *PLoS Comput Biol* **9**, e1002861 (2013).
46. Aziz, M. F., Caetano-Anollés, K. & Caetano-Anollés, G. The early history and emergence of molecular functions and modular scale-free network behavior. *Sci. Rep.* **6**, (2016).
47. Wang, M. *et al.* A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation. *Mol. Biol. Evol.* **28**, 567–582 (2011).
48. Caetano-Anollés, D., Kim, K. M., Mittenthal, J. E. & Caetano-Anollés, G. Proteome evolution and the metabolic origins of translation and cellular life. *J. Mol. Evol.* **72**, 14–33 (2011).
49. Mittenthal, J. E., Caetano-Anollés, D. & Caetano-Anollés, G. Biphasic patterns of diversification and the emergence of modules. *Front. Genet.* **3**, 147 (2012).
50. Aziz, M. F. *et al.* Stress induces biphasic-rewiring and modularization patterns in the metabolomic networks of *Escherichia coli*. *IEEE Intl. Conf. Bioinf. Biomed.* 593–597 (2012) doi:10.1109/BIBM.2012.6392626.
51. MacDougall, M. H. Simulating computer systems: Techniques and tools. (1987).
52. Delaney, W. & Vaccari, E. Dynamic models and discrete event simulation. (1989).
53. Pidd, M. Computer simulation in management science. *JOURNAL-OPERATIONAL Res. Soc.* **57**, 327 (2006).
54. Caetano-Anollés, G., Kim, H. S. & Mittenthal, J. E. The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proc. Natl. Acad. Sci.* **104**, 9358 (2007).
55. Caetano-Anollés, G., Kim, K. M. & Caetano-Anollés, D. The phylogenomic roots of modern biochemistry: Origins of proteins, cofactors and protein biosynthesis. *J. Mol. Evol.* **74**, 1–34 (2012).

56. Caetano-Anollés, K. & Caetano-Anollés, G. Structural phylogenomics reveals gradual evolutionary replacement of abiotic chemistries by protein enzymes in purine metabolism. *PLoS One* **8**, e59300 (2013).
57. Caetano-Anollés, G. *et al.* The origin and evolution of modern metabolism. *Int. J. Biochem. Cell Biol.* **41**, 285–297 (2009).
58. Barabási, A. L. & Albert, R. Emergence of scaling in random networks. *Science* (80-.). **286**, 509 (1999).
59. Pang, T. Y. & Maslov, S. Universal distribution of component frequencies in biological and technological systems. *Proc. Natl. Acad. Sci.* **110**, 6235–6239 (2013).
60. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A. L. Hierarchical organization of modularity in metabolic networks. *Science* (80-.). **297**, 1551–1555 (2002).
61. Erdős, P. & Rényi, A. Connectivity of random nets. *Publ. Math. Inst. Hungarian Acad. Sci.* **5**, 17–61 (1960).
62. Bollobas, B. *Random Graphs*. (Academic Press, London, 1985).
63. Strogatz, S. H. Exploring complex networks. *Nature* **410**, 268–276 (2001).
64. Mughal, F. & Caetano-Anollés, G. MANET 3.0: Hierarchy and modularity in evolving metabolic networks. *PLoS One* **14**, e0224201 (2019).
65. Newman, M. E. J. & Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 26113 (2004).
66. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A. L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
67. Overbeek, R. *et al.* WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* **28**, 123–125 (2000).
68. Wagner, A. & Fell, D. A. The small world inside large metabolic networks. *Proc. R. Soc. London. Ser. B Biol. Sci.* **268**, 1803–1810 (2001).
69. Wasserman, S. & Faust, K. *Social network analysis: Methods and applications.* **8**, (1994).
70. Barrat, A., Barthélemy, M., Pastor-Satorras, R. & Vespignani, A. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 3747–3752 (2004).
71. Watts, D. J. & Strogatz, S. H. Collective dynamics of ‘small-world’ networks. *Nature* **393**, 440–442 (1998).
72. Albert, R. & Barabási, A. L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47 (2002).
73. Newman, M. E. J., Strogatz, S. H. & Watts, D. J. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 26118 (2001).

74. Caetano-Anollés, G., Wang, M. & Caetano-Anollés, D. Structural phylogenomics retrodicts the origin of the genetic code and uncovers the evolutionary impact of protein flexibility. *PLoS One* **8**, e72225 (2013).
75. Tal, G., Boca, S. M., Mittenthal, J. & Caetano-Anollés, G. A dynamic model for the evolution of protein structure. *J. Mol. Evol.* **82**, 230–243 (2016).
76. Mrvar, A. & Batagelj, V. Analysis and visualization of large networks with program package Pajek. *Complex Adapt. Syst. Model.* **4**, 1–8 (2016).
77. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Syst.* **1695**, 1–9 (2006).
78. Van Eck, N. J. & Waltman, L. VOS: a new method for visualizing similarities between objects. in *Advances in Data Analysis: Proceedings of the 30th Annual Conference of the German Classification Society* 299–306 (Heidelberg: Springer Verlag, 2007).
79. Waltman, L., van Eck, N. J. & Noyons, E. C. M. A unified approach to mapping and clustering of bibliometric networks. *J. Informetr.* **4**, 629–635 (2010).
80. Kamada, T. & Kawai, S. An algorithm for drawing general undirected graphs. *Inf. Process. Lett.* **31**, 7–15 (1989).
81. Ihaka, R. & Gentleman, R. R: A language for data analysis and graphics. *J. Comput. Graph. Stat.* **5**, 299–314 (1996).
82. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. (2014).
83. PHP-Group & others. PHP: Hypertext PreProcessor. *Internet* <http://www.php.net> (2012).
84. Newman, M. E. J. Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* **46**, 323–351 (2005).
85. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM Rev.* **51**, 661–703 (2009).
86. Clauset, A., Newman, M. E. J. & Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **70**, 66111 (2004).
87. Borg, I. & Groenen, P. Modern multidimensional scaling: theory and applications. *J. Educ. Meas.* **40**, 277–280 (2003).
88. Murtagh, F. & Legendre, P. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *J. Classif.* **31**, 274–295 (2014).
89. Bartels, R. The rank version of von Neumann's ratio test for randomness. *J. Am. Stat. Assoc.* **77**, 40–46 (1982).
90. Erdős, P. & Rényi, A. On random graphs, I. *Publ. Math.* **6**, 290–297 (1959).

91. Chung, F. R. K., Erdős, P. & Spencer, J. On the decomposition of graphs into complete bipartite subgraphs. in *Studies in pure mathematics, To the memory of Paul Turán* (eds. Erdős, P., Alpar, L., Halasz, G. & Saekőzy, A.) 95–101 (Budapest: Birkhäuser Verlag, 1983).
92. Diestel, R. *Graph theory. Graduate Texts in Mathematics*. vol. 173 (Heidelberg: Springer-Verlag, 2010).
93. Winter, C., Henschel, A., Kim, W. K. & Schroeder, M. SCOPPI: a structural classification of protein–protein interfaces. *Nucleic Acids Res.* **34**, D310–D314 (2006).
94. Teichmann, M., Dumay-Odelot, H. & Fribourg, S. Structural and functional aspects of winged-helix domains at the core of transcription initiation complexes. *Transcription* **3**, 2–7 (2012).
95. Berezovsky, I. N. & Trifonov, E. N. Van der Waals locks: loop-n-lock structure of globular proteins. *J. Mol. Biol.* **307**, 1419–1426 (2001).
96. Aharonovsky, E. & Trifonov, E. N. Protein sequence modules. *J. Biomol. Struct. Dyn.* **23**, 237–242 (2005).
97. Goncarenko, A. & Berezovsky, I. N. Prototypes of elementary functional loops unravel evolutionary connections between protein functions. *Bioinformatics* **26**, i497–i503 (2010).
98. Kim, H. S., Mittenthal, J. E. & Caetano-Anollés, G. Widespread recruitment of ancient domain structures in modern enzymes during metabolic evolution. *J. Integr. Bioinform.* **10**, 214 (2013).
99. Teichmann, S. A., Chothia, C. & Gerstein, M. Advances in structural genomics. *Curr. Opin. Struct. Biol.* **9**, 390–399 (1999).
100. Taylor, B. L. & Zhulin, I. B. PAS domains: internal sensors of oxygen, redox potential, and light. *Microbiol. Mol. Biol. Rev.* **63**, 479–506 (1999).
101. Apic, G., Gough, J. & Teichmann, S. A. An insight into domain combinations. *Bioinformatics* **17**, S83 (2001).
102. Tanaka, R. Scale-rich metabolic networks. *Phys. Rev. Lett.* **94**, 168101 (2005).
103. Li, L., Alderson, D., Tanaka, R., Doyle, J. C. & Willinger, W. Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications (Extended Version). *Arxiv Prepr. cond-mat/0501169* (2005).

CHAPTER 3: UNTANGLING THE BIRTH OF STRUCTURAL DOMAINS BY RECRUITMENT OF PROTEIN LOOPS WITH CHRONOLOGIES, NETWORKS, AND *AB INITIO* STRUCTURAL PREDICTION

3.1 Abstract

The structures and functions of proteins are embedded into the loop scaffolds of protein domains. Their evolution remains mysterious. Here, we use a bipartite network approach to study the history of loop prototypes, elementary units that combine with each other to form domain structures in protein evolution. Loops were classified based on structural geometry and conformation using ArchDB. Structural domains were defined at fold family level of SCOP classification and their evolution tracked with phylogenomic chronologies. The resulting bi-modal ‘elementary functionome’ (EF) network and its loop and domain graph projections described evolutionary ‘waterfalls’ emphasizing the emergence of primordial functions. Timed waterfalls revealed historical sharing of ancient functional loops by domains in two primordial waves of functional innovation that involved founder ‘*p*-loop’ and ‘winged helix’ domains. We also uncovered a dual hourglass recruitment pattern of embedding of modular motifs in structures, exchanging ‘preferential’ cooption properties between ancient loops and emerging domains. Notably, we found the emergence of molecular functions to be modernistic and ongoing. Networks of loop motifs and domain structures expanded metabolic pathways before translation functionalities while simultaneously fostering hierarchical modularity in network evolution. The origin and evolution of domains is backed by strong genomic and structural evidence, which can be used to model the gradual evolution of proteins, as we illustrate with AlphFold2-modelled P-loop hydrolases.

3.2 Introduction

“... *I arrive now at the ineffable core of my story. And here begins my despair as a writer. All language is a set of symbols whose use among its speakers assumes a shared past. How, then, can I translate into words the limitless Aleph, which my floundering mind can scarcely encompass?...*”— Jorge Luis Borges, *The Aleph and Other Stories*

The protein world is both structured and functionally complex. Its emergence and history merits exploration. In protein evolution, structural, functional, and evolutionary units of proteins called *structural domains* must emerge from prior structural states by fulfilling the evolutionary principle of spatiotemporal continuity (*‘lex continui’* of Leibnitz). These prior states likely involve a combinatorial origami of dipeptides capable of forming flexible protein loop structures, which led to coevolutionary interactions with nucleic acid cofactors and the rise of genetics (1) (Caetano-Anollés et al., 2013). Prior states may also involve evolutionary stable and functionally relevant loop intermediates capable of giving rise to the enormous diversity of protein domains that exist in nature (2,3)(Trifonov and Frenkel, 2009; Sobolevsky et al., 2013). We here explore such scenario of emergence with structural phylogenomics and evolving networks.

Loops define a diverse group of *supersecondary* building blocks made of helix, strand, turn and coil segments that are generally ~25 to 30 amino acid residues long, much smaller than the size of an average compact domain (~100 amino acid residues in length)(4,5)(Söding and Lupas, 2003; Papaleo *et al.*, 2016). Loop structures embody non-regular (aperiodic) loop regions spanning ‘helical’ and ‘sheet’ structural components (6)(Leszczynski and Rose, 1986), which direct the polypeptide chain in space and are often functionally important. Supersecondary ‘closed loop’ structures collapse into extended flexible or rigid loop-shaped primordial intermediate

conformations stabilized by van der Waals locks (7,8)(Berezovsky and Trifonov, 2001; Aharonovsky and Trifonov, 2005). Loop prototypes are universally present in structures and are regarded as modern determinants of molecular function (9,10)(Berezovsky et al., 2000; Goncarenko and Berezovsky, 2010). While their biophysical properties may constrain their evolution, studies identified evolutionarily conserved loop prototypes that were likely responsible for the early rise of molecular functions in protein evolution. A first group of ‘elementary functional loop’ (EFL) prototypes combine with others to form active sites that bind cofactors and exert molecular functions (10,11,12,13)(Goncarenko and Berezovsky, 2010, 2012, 2015; Berezovsky *et al.*, 2017). These EFLs were obtained by iterative derivation of sequence profiles from protein coding sequences in complete proteomes using a scoring function that weights profile positions according to their information content (10)(Goncarenko and Berezovsky, 2010). Distant evolutionary connections of EFLs revealed patterns of motif reuse in archaeal proteins (11)(Goncarenko and Berezovsky, 2012). A chronology of bipartite networks linked domains to EFLs (and their projections) and showed that the multifunctional α - β - α layered design typical of P-loop and Rossmann-like sandwich structures was primordial (14)(Aziz *et al.*, 2016). The networks also showed EFL recruitment events occurring throughout the 3.8 Gy history of proteins, suggesting the origin of novel domains is an ongoing process. In contrast with EFLs, a second group of highly repeated non-combinable loop structures present in popular folds indicate remnants of an ancient peptide ‘vocabulary’ that formed folded polypeptides during a primordial RNA-peptide world (15)(Alva *et al.*, 2015). Using machine learning methods, biphasic patterns in probability distributions highlighted high-scoring subdomain-sized fragments that were unified by <30% sequence similarities. These fragments were aligned into folded loop structures that were 9-39 residue long, embedding helix-turn-helix, helix-hairpin-helix, ribosomal protein, P-

loop/dinucleotide-binding β - α - β (catalytic), and metal ion/iron-sulfur cluster (binding function) motifs. A third general type of supersecondary structural motifs involve widely reused, contiguous, and non-overlapping segments with longer lengths varying from 35 to 200 amino acids (16,17)(Nepomnyachiy *et al.*, 2014, 2017). These so-called ‘themes’ have been used to build networks of domains and motifs linked by motif reuse in domains (16)(Nepomnyachiy *et al.*, 2014), which interestingly increases with decreasing theme length following a power law, (17)(Nepomnyachiy *et al.*, 2017). The power law indicates a significantly biased distribution of themes in proteins. All three approaches characterize (i) supersecondary motifs by sequence and/or structure similarities, not necessarily carrying any evolutionary relationship; (ii) motif recurrence across proteins driven by biological function; and (iii) complex patterns showcasing an interplay of divergent vs. convergent evolution driven by rearrangements, duplications, and divergences. EFLs, loops and themes therefore represent ancient building blocks that are evolutionarily conserved.

We previously used a phylogenomic framework for reconstruction of evolutionary timelines of molecular accretion built from the sequence and structure of thousands of nucleic acid molecules and millions of protein sequences encoded in thousands of genomes [reviewed in (18,19) Caetano-Anollés *et al.*, 2009, 2021]. Chronologies made protein evolution explicit by tracing the gradual appearance of domain structures (20,21)(Caetano-Anollés and Caetano-Anollés, 2003; Edwards *et al.*, 2013), their combinatorial rearrangement in proteins (22)(Wang and Caetano-Anollés, 2009), and their associated molecular functions (23-24) (Kim and Caetano-Anollés, 2010; Koç and Caetano-Anollés, 2017), along with chemical and biophysical properties of significance. For instance, tracing chemical mechanisms in enzymatic reactions uncovered the natural history of biocatalysis (25)(Nath *et al.*, 2014). Similarly, tracing contact order (i.e., the average relative

distance of amino acid contacts in the tertiary structure of proteins), which correlates to flexibility, showed that folding speed is optimized and increases during a first phase of protein evolution (26)(Debès *et al.*, 2013).

We focus on loops sourced from the ArchDB database (27)(Bonet *et al.*, 2014). ArchDB provides an exhaustive classification of loop structures into loop prototypes (supersecondary motifs) based on both a Density Search (DS) clustering algorithm and a graph-based Markov clustering (MCL) algorithm, both of which are structural alignment (RMSD)-independent. Both implementations explore a multidimensional feature space defined by the number of amino acid residues (length) of aperiodic structure, bracing secondary structures, and the conformation and geometry of loop structures. Our analysis makes use of the more stringent DS ‘mode-seeking’ classification method, which detects regions of feature space with high loop density organized around centroids. The method fixes the ‘length’ of loops and enlarges the coverage of clustered groups. We use evolving networks to trace the coevolutionary history of loop prototypes (herein often referred to as ‘loops’) and protein structural domains defined at the fold family (domain) level of the SCOP domain classification (28)(Murzin *et al.*, 1995) (herein referred to simply as ‘domains’). Our evolving network reveals remarkable patterns of emergence at molecular level, describing how loops of ancient and more recent origin combine to form domain structures in protein evolution. The method allows to model the emergence of the folded structure of domains using *ab initio* structural prediction.

3.3 Results and Discussion

3.3.1 Reconstructing the history of an ‘elementary functionome’ of loop structures

We traced the times of origin (evolutionary age) of domain structures over a bimodal graph-theoretic representation of domains and their associated loop prototypes. Later, we decomposed the bipartite representation into monomodal network projections using mathematical properties of finite graphs (29)(Chung, Erdős and Spencer, 1983). The general strategy is described in Fig. 3.1. The data pipeline involved the survey of domains and loops, their mapping to each other with bipartite networks, the phylogenomic reconstruction of a chronology (series of time events) derived from a phylogeny that describes the evolution of domains, and the unfolding of a time series of networks describing recruitment patterns and evolution of an ‘elementary functionome’ (EF) of loop structures that are modular (Fig. 3.1A).

Loops were classified into prototypes using the exhaustive classification scheme of ArchDB (27)(Bonet *et al.*, 2014) based on structural geometry and conformation (Fig. 3.1B). Each ArchDB loop structure is a region of a PDB entry that associates with one ArchDB-classified loop prototype, which in turn makes up the structure of one or many domains. In our study, prototypes were identified using DS filtering of domains to loop mappings with e-value <0.001 (27)(Bonet *et al.*, 2014). They were named according to clustering method used (DS), ‘type’ of bracing secondary structures (HH, α -helix– α -helix; HE, α -helix– β -strand; EH, β -strand– α -helix; HG, α -helix– 3_{10} -helix; GH, 3_{10} -helix– α -helix; GG, 3_{10} -helix– 3_{10} -helix; EG, β -strand– 3_{10} -helix; GE, 3_{10} -helix– β -strand; BN, β - β hairpin; and BK, β - β link), length of the aperiodic loop region between secondary structures, class (same conformation) and subclass (common geometry), in that order. For example, the ancient DS.HE.3.1.1 prototype present in ancient NAD(P)-binding Rossmann-

fold domains has α -helix and β -strand bracing structures (HE), a 3 residue-long loop region, and the most populated conformation and geometry classes (ranked 1) (Fig. 3.1B). Structural domains defined at domain level were named using SCOP *concise classification strings* (ccs). For example, the tyrosine-dependent oxidoreductase domain that holds the ancient DS.HE.3.1.1 prototype has a ccs of c.2.1.2 typical of Rossmann folds. The ‘times of origin’ of domains were directly derived from a most parsimonious phylogenomic tree of domain structures generated from a census of domain abundance in 8,127 proteomes (Fig. 3.1C). A heat diagram of the phylogenomic data matrix already reveals tantalizing abundance patterns differentiating the proteomes of Archaea, Bacteria, Eukarya, and viruses. The highly unbalanced tree of domains permits to establish times of origin of domains in a relative 0-to-1 scale of node distance units (*nd*). A molecular ‘clock of folds’ derived from calibration points of protein domain structures defined associated with microfossil, fossil and biogeochemical evidence [including molecular, physiological, paleontological, and geochemical markers and first appearance of clade-specific domains; first described in (30) Wang et al. 2011] was used to convert relative *nd* ages of domain domains into geological time in billions of years (Gy) (Fig. 3.1C). As expected (31,32)(Caetano-Anollés et al., 2012; Caetano-Anollés & Caetano-Anollés, 2013), *nd* values of the most ancient domains in fold superfamilies were strongly correlated with geological time (Fig. 3.1D). The chronology allowed to transfer the times of origin of domains to loops, imposing time directionality on network links (making them arcs with arrows pointing from older to younger nodes) and allowing construction of time series of networks that are growing in evolutionary time using methodologies developed in Aziz et al. (14)(2016). Since ancestral loops are recruited into growing structures of domains to perform modern functions, their time of origin were borrowed from the most ancestral linked domains or from the second oldest domains when multiple domains shared loop structures.

The EF network is a bipartite graph with two segregate sets of vertices (nodes), one representing loops (circles) and the other domains (rhomboids) (Fig. 3.1E). Bipartite graphs project connectivity that is shared within each set of nodes (33)(Diestel, 2010). Inter-links (undirected edges) between nodes of the two groups become the basis for intra-links between nodes of each group, with their respective connectivity being correlated. Thus, links in the domain projection describe how domains share loops in their structural makeup, and conversely, links in the loop projection describe how loops combine to form structures around active sites in domains. Since loops host molecular functions (5)(Papaleo *et al.*, 2016) and domains represent bona fide structural, functional and evolutionarily conserved units of proteins (28)(Murzin *et al.*, 1995), the time series of EF networks and their projections described how domains recruited molecular functions in protein evolution.

The progress of a timed network system can be modeled as a discrete sequence of events (network growth at temporal intervals) using Discrete Event Simulation (DES)(34-36)(Delaney and Vaccari, 1989; MacDougall, 1987; Pidd, 2006). DES is a typical computer-based modeling tool used to analyze complex network behavior and how it affects the interacting entities on consecutive events. Under DES, a system stays consistent between events. Consequently, time behaves as a *step function*, flowing from event to event in a discretized manner. Here we borrow the DES rationale and map discrete evolutionary ‘time steps’ (intervals) to the growing structure of the undirected and unweighted EF network and its directed weighted projections (Fig. 3.1E). This is further illustrated with a toy example in Supplementary Fig. S1.

3.3.2 A frustrated and ongoing history of modular and non-modular loop recruitment

We clustered ArchDB loop structures into 9,650 loop prototypes, which in turn mapped to the 3,892 domain families of our phylogenomic timeline. This initial set of loops was then filtered at e-value <0.001. This clustered 88,321 loops structures into 7,078 loop prototypes, establishing many-to-many mappings to 2,447 domains. Filtering minimized false positives while detecting reliable structural and functional associations at statistically significant levels. The 7,078 loop prototypes were divided into three subsets according to how they associated with domains across the evolutionary timeline (Fig. 3.2A). A subset of 5,125 '*non-modular*' (NM) prototypes mapped uniquely to individual domains of a same age belonging to a set of 1,965 domains. In contrast, a subset of 1,937 '*modular*' (M) prototypes mapped to more than one domain out of 1,442 domains with times of origin spread throughout the timeline. These prototypes acted as modular units of structural, functional and evolutionary significance. They represented the most abundant, widely distributed, and interesting loops of this study. Finally, a small subset of 16 '*modular*' (contemporaneous) (M') prototypes involved associations with more than one domain that occurred within individual time events. While we define modules as sets of integrated (coordinated) parts that cooperate to perform a task and interact more extensively with each other than with other parts and modules of a system, we recognize modules by the property of 'modularity', the degree to which parts of a system can be separated and rearranged in different contexts.

A Venn diagram of domains mapping to the three groups of prototypes revealed three Venn groups of especial interest, domains mapping to only M loops (313 domains), domains mapping to only NM loops (827), and domains mapping to both M and NM loops (1,109) (Fig. 3.2A). The fact that a significant number of domains are recruiting both M and NM loops suggest their makeup is

shaped by two recruitment mechanisms, one stochastic and the other evolutionary. It is likely that NM loops are stochastically drawn into domains by local genome rearrangement activities but are never coopted by other domains because they fail to add significant protein functionalities. In contrast, M and M' loops behave as true evolutionary units capable of distributing structural and functional novelties to many domains along the timeline. Their evolutionary recruitment enables a combinatorial origami of modular structural motifs benefiting protein evolution. Fig. 3.2B traces the three Venn groups of domains and the three groups of loop prototypes along the evolutionary chronology. Remarkably, domains mapping to both M and NM loops accumulated earlier than domains mapping only to either M or NM domains along the chronology. In addition, the presumed more stochastic NM loops also accumulated earlier than the M and M' loops in evolution. Our observations are compatible with highly dynamic views of protein organization [e.g. (37,38) Tawfik, 2010; Turoverov et al., 2019] or the existence of molecular discriminating Maxwell demons that dissipate energy and information (39)(Boël et al., 2019). These views foster molecular systems that are frustrated by 'messiness' in the form of stochastic noise, heterogeneity, infidelity, and variation, as nicely exemplified by the existence of intrinsic disorder embedded in the structural makeup of proteins.

3.3.3 Time event 'waterfall' networks uncover the birth of domains in protein evolution

To further dissect the different recruitment strategies, we constructed two bipartite networks, one describing the evolutionary recruitment of NM loops and the other describing the recruitment of M loops into the structure of domains. The evolving bipartite network that links NM loops to domains uncovered 'horizontal' recruitments occurring throughout protein evolution, always restricted to individual time events (Fig.3A). Note that meaningful network projections cannot be produced because there are no links other than the horizontal single loop-to-domain mappings. A

multitude of ‘temporal’ recruitment waves in the form of ‘ripples’ were however evident throughout the timeline, with at least 5 significant ripples occurring between 3.8 and 2.7 Gya and two subsequent (major and distinctive) ripples occurring ~2.5 Gya and ~2 Gya. These series of small waves embody a multiplicity of subnetworks unfolding in time, which is in sharp contrast with waves that involve individual subnetworks with single origins, which we will later discuss. The recurrent patterns of the waves in these ripples clearly show that the stochastic mechanisms of cooption we propose are ongoing. In sharp contrast, the evolving bipartite network that links M loops and domains shows ‘vertical’ recruitments occurring between the 1,937 modular loop prototypes and 1,442 domains throughout the timeline (Fig. 3.3B). Visual inspection of the network showed waves of co-option, some of which matched the major ~2.5 Gya and ~2 Gya ripples of the non-modular network and involved pervasive recruitment of older loops. This network is the most significant because it explains how loop recruitments have shaped the structure of domains in the protein world. It is truly an EF network, which can be fully dissected into loop and domain projections. Given its evolutionary centrality, our focus will shift to this network.

The bipartite EF network of Fig. 3.3B embodies an undirected graph of 2,546 links with a *network density* (actual/possible number of links) of 0.0009 [2546/(1442×1937)] and a *node average degree* (links per node) of 1.507 (± 0.018), i.e. loop components had approximately less than 2 interlinks on average. Visual inspection revealed the network was well clustered. The Visualization of Similarity (VOS) clustering method (40,41)(Van Eck and Waltman, 2007; Waltman et al., 2010) uncovered 889 communities (also known as modules) with a high *modularity index* of 0.996. The adaptive events of the EF network were made vivid by color coding and arranging the component nodes by age in a top-down bimodal layout that followed the evolutionary timeline of domain structures (Fig. 3.3B). The node sizes were made proportional to

node connectivity, measured by weighted degree, highlighting the hub-like behavior of the network structure. In order to better visualize evolutionary network patterns, the VOS clusters (comprising of hubs and their neighbors) were spread horizontally using the energy-optimized Kamada-Kawai (42)(Kamada and Kawai, 1989) ‘free’ and ‘optimize inside clusters only’ methods (Fig. 3.3C). The resulting ‘waterfall’ network layouts vividly illustrate functional recruitment across and down the events of the waterfall that is responsible for how loop modules make up domains in evolution. Consequently, EF network projections were also visualized in waterfall layouts (Fig. 3.4). The connectivity of these monomodal networks was made evolutionarily explicit by giving a direction to the intra connecting nodes with *arcs*. The resulting loop and domain directed networks had 2,024 and 1,005 arcs each, with *network densities* of 0.00054 $[2024/(1937 \times (1937 - 1))]$ and 0.00048 $[1005/(1442 \times (1442 - 1))]$ and total *node average degrees* of 2.104 (± 0.061) and 1.415 (± 0.067). In other words, ~2 domains and ~1.5 loops were shared on average, respectively. Both the loop and domain monomodal networks also showed significantly high community structure with 879 and 882 clusters each, and modularity indices of 0.994 and 0.990, respectively. The number of outward (outdegree) and inward (indegree) directed connections or arcs specified domain and loop nodes as ‘donors’ (sources) or ‘acceptors’ (sinks) of shared functional loops and domains, respectively. The horizontal and vertical scale of node symbols were made proportional to the weighted outdegree and indegree, respectively. This made the visualization of hubs explicit, e.g., a transition from wide to tall symbols along the flow of time indicated source-sink origination transformations of molecular functions. Overall, these transitions expressed an expected increase in the probability of co-opting older loops and domains with time, but also a surprisingly continual recruitment process operating among recent molecular forms.

3.3.4 The evolutionary emergence of scale-free properties in the EF network

Growing networks that follow the preferential attachment principle have large, highly connected hubs that attract increasingly more links in a ‘rich-get-richer’ fashion. In these highly inhomogeneous networks, which are ubiquitous in biology, the probability $P(k)$ of a node linked to k nodes (i.e., the fraction of nodes with k links or k -neighbors) decays as a power law, $P(k) \sim k^{-\gamma}$, devoid of a characteristic scale. ‘Scale-free’ networks in general have regression exponents $\gamma = 2.1$ – 2.4 , with a typical heavy-tailed distribution in which very few nodes have high connectivity (43)(Strogatz, 2001). For metabolic reaction networks, $\gamma = 2.2$ in all organisms (e.g., (44)(Jeong *et al.*, 2000). To test if the evolving EF network and its projections had degree distributions that tended to follow the scale-free properties, we studied the cumulative connectivity (of links or arcs) along the chronology of the growing networks and tested power law behavior with appropriate statistics (Fig. 3.5A; Supplementary Fig. S2). Remarkably, we noticed a power law tendency, along with associated ‘scale free’ generative models, was an emergent property in the EF network, but not in its projections.

A few statistics failed to reject power law properties in connectivity of the most ancient loops and domains very early in evolution ($nd \sim 0$ – 0.02), of domains but not loops during the $nd \sim 0.02$ – 0.04 interval, then of loops but not domains during $nd \sim 0.04$ – 0.07 , and finally of loops but not domains of the EF bipartite network throughout the rest of the timeline (Fig. 3.5A; Supplementary Fig. S2). The most prominent indicator of failure to reject was the Kolmogorov-Smirnov (KS) statistical test of power law fit (45,46)(Clauset *et al.*, 2009; Newman, 2005) (Fig. 3.5A). High p-values of the KS test (≥ 0.05) and low values of the KS fit statistic (≤ 0.10) failed to reject network data being drawn from the fitted power-law distribution. Similarly, the exponent of the fitted power law distribution (α), which is $\alpha > 1$ when the assumption of probability of power law fit $P(X = x^{-\alpha})$ is

met, was incidentally higher, for example, for later degree distributions of loop components of the growing EF network. These patterns are indicative of power-law decay. However, the log-likelihoods of the fitted power law parameters gradually deviate towards values lower than zero for most network events of the timeline, making typical power law distributions less likely. The statistical analyses of the timeline of growing EF networks therefore reveal a surprising property of power law emergence and transfer. While the early-evolved loop ‘prior form’ component of the bipartite network shared power law behavior with the domain component, these ancient components went through two cycles of an exchange of scale-free properties, from domains to loops and then from loops to domains, as molecular functions developed in protein evolution (Fig. 3.5A; Supplementary Fig. S2).

Log-linear regression models overlapping the power law curves showed that the coefficient of power law decay γ for the domain portion of the EF network varies in the beginning from a linear scale ($\gamma=1.000$, $nd\sim 0.01$) to a moderate ($\gamma=1.585$, $nd\sim 0.02$), then a strong ($\gamma=2.000$, $nd\sim 0.04$), and then again a moderate ($\gamma=1.404$, $nd\sim 0.07$) power law scale, while passing through another linear phase ($\gamma=1.000$, $nd\sim 0.06$) before continuing to increase to a very strong power law scale on average ($\gamma=2.456\pm 0.034$ since $nd\sim 0.31$, with max $\gamma=2.915$ at $nd\sim 0.84$). For the loop portion, however, while γ also starts with a linear scale ($\gamma=1.000$, $nd\sim 0.01$), it varies somewhat symmetrically to that for domains, in the beginning with a shift to an antilinear scale ($\gamma=-1.000$, $nd\sim 0.02$) before fluctuating between a strong ($\gamma=2.265$, $nd\sim 0.04$), moderate ($\gamma\sim 1.807$, $nd\sim 0.05$), and then very strong ($\gamma=2.459$, $nd\sim 0.07$) power law scale, subsequently climbing to a highly strong power law scale on average with a slight lag ($\gamma=2.877\pm 0.046$ since $nd\sim 0.33$, with max $\gamma=3.450$ at $nd\sim 0.85$). Initial fluctuations were evident well before the first wave of the $\sim 2,4$ Gya ripple we mentioned above. The coefficient of determination (R^2) of $\sim 95\%$ supports the linear models for both the

domain and loop components. Thus, the extent of preferential attachment in protein evolution is more robust than that reported for metabolic networks (47)(Ravasz *et al.*, 2002). Remarkably, loop ($\gamma=2.270\pm0.030$) and domain ($\gamma=1.865\pm0.038$) network projections maintain only a strong and moderate power law scale. The extraordinary observation of a dual yin-and-yang like power law transfer between loop and domain components may be indicative of a continuing global scaling phenomenon in a biphasic emergence of biological modules (48)(Mittenthal, Caetano-Anollés and Caetano-Anollés, 2012), which we now explain. Note that we define the power law arbitrary scales used above as follows: moderate $\geq \sim 1.3$, strong $\geq \sim 2.0$, very strong $\geq \sim 2.3$, highly strong $\geq \sim 3.0$.

3.3.5 The rise of hierarchical modularity

Networks become modular when their nodes connect to each other within bounds of a community (module) (49)(Newman & Girvan, 2004). Modularity counteracts the scale-free property of biological networks by equalizing the degree distribution of nodes in the modules (44,50)(Jeong *et al.*, 2000; Wagner & Fell, 2001). However, these opposing properties reconcile when modules are integrated hierarchically (47)(Ravasz *et al.*, 2002). A primary measurement of modularity is the *average clustering coefficient* (C), a ratio of triangles (graph cycles of length 3) to the connected triads in the graph, averaged over all nodes, while ignoring the direction and weights of the edges (47,51,52)(Wasserman and Faust, 1994; Ravasz *et al.*, 2002; Barrat *et al.*, 2004). Since C for the bipartite EF network was not meaningful due to absence of triangles, modular organization was investigated through its projections (Fig. 3.5B; Supplementary Fig. S3). The domain and loop networks exhibit C values of $\sim 0.805 (\pm 0.0066)$ and $\sim 0.893 (\pm 0.0025)$, respectively, significantly higher than ~ 0.6 reported for metabolic networks (45,50,53)(Watts and Strogatz, 1998; Wagner and Fell, 2001; Ravasz *et al.*, 2002). The elevated C of EF network projections suggests that densely intra-connected modules of loop prototypes and domain

structures are integrated by few inter-modular links. Thus, the EF network has a highly consolidated modular structure.

A notable property of C is its sharp decline with network size N for scale-free models (54)(Albert and Barabási, 2002), as $N^{-0.75}$, contrary to highly modular networks that are independent of N (47)(Ravasz *et al.*, 2002). For the domain and loop networks, C regressed with N as $N^{-0.00019}$ and $N^{-0.000033}$, and with age nd of the networks as $nd^{-0.35}$ and $nd^{-0.091}$, respectively (Fig. 3.5B; Supplementary Fig. S3), confirming the modular structure of the evolving networks. The smaller exponents suggest the increased ‘granularity’ of the modular makeup of the loop network compared to that of the domain network, supporting earlier observations that lower levels of organization in bipartite networks of metabolism were more granular and cohesive (55)(Mughal and Caetano-Anollés, 2019). As expected, C of strictly power-law (Barabási) reference controls were zero (56)(Newman *et al.*, 2001). The timeline of growing domain and loop networks revealed anticorrelated trends of modularity and scale-free properties. For example, the C of domain and loop networks show two initial cycles of fall and rise in modularity (a drop from 1.000 to ~ 0.889 and ~ 0.950 at $nd \sim 0.112$, rise to ~ 0.956 and ~ 0.894 at $nd \sim 0.258$, and then a dip to ~ 0.830 and ~ 0.892 at $nd \sim 0.356$, respectively), followed by a plateau to ~ 0.73 and ~ 0.88 at $nd \sim 0.635$, respectively. These patterns match the power law trends, as indicated by KS fit indegree statistic, which manifest slightly earlier than the corresponding modularity phases and reject power law behavior in initial phases (a rise from 0.000 to ~ 0.2 each at $nd \sim 0.077$, fall to ~ 0.1 each at $nd \sim 0.146$, and then a peak to ~ 0.2 and ~ 0.12 at $nd \sim 0.202$, respectively), before plateauing to ~ 0.06 and ~ 0.1 at $nd \sim 0.339$, respectively (Fig. 3.5B; Supplementary Figs. S2 and S3). Since these counteracting trends of modularity that follow those of preferential attachment with lags of $\Delta nd \sim [0.035, 0.112, 0.154, 0.296]$ in the domain and loop networks must be reflective of an emergent power law

behavior of a modular EF bipartite network, our findings highlight a notable conjecture. Transfer of scale-free properties from loop to domain components of the EF involves a generative cycle of modular and hierarchical structure of interacting loop prototypes and domain structures.

To verify this conjecture, we also analyzed the dynamics of three measures of network modularity along the timeline of growing networks. The Newman-Girvan (*NG*) algorithm iteratively calculates edge ‘betweenness’, i.e., the maximum number of shortest paths running through an edge, while systematically removing edges scoring high in inter-community centrality (49)(Newman and Girvan, 2004). This perpetual removal uncovers the modular structure of a network measured by the *NG* index. *NG* partitioned by age (NG_{age}) ranges from -1 to 1 , with positive values indicating modular connectivity within events, while negative values indicating otherwise. *NG* partitioned by VOS (NG_{vos}) describes the cohesiveness of VOS memberships (40,41)(Van Eck and Waltman, 2007; Waltman, van Eck and Noyons, 2010). The Fast Greedy Community (*FGC*) detection algorithm provides a hierarchical perspective of agglomerative community structure (57)(Clauset et al., 2004). Noticeably, NG_{vos} and *FGC* indices produced mirrored patterns of growth of modular cohesiveness and agglomerative structure by age for all growing networks. Conversely, NG_{age} indicated divergent progress towards age-associativity among the bipartite EF network and its domain and loop network projections. All three networks showed age-independent origins with $NG_{age} \leq -0.3$, -0.25 and -0.5 until $nd \sim 0.034$, ~ 0.073 and ~ 0.039 , respectively. Loop and domain networks then parsimoniously progress towards age-associativity with an early rise of NG_{age} to ~ -0.2 and then to ~ -0.1 , at $nd \sim 0.039$ and ~ 0.056 , respectively, in the EF network, and at $nd \sim 0.077$ and ~ 0.112 , respectively, in the domain projection, before plateauing out to ~ -0.01 at $nd \sim 0.545$ in EF and to ~ -0.006 at $nd \sim 0.575$ in the domain network. However, the loop projection became aggressively age-associative along the

early timeline, with an initial spike of NG_{age} from -0.5 to ~ -0.07 at $nd \sim 0.43$ followed by a gradual rise to ~ 0.29 at $nd \sim 0.18$, before plateauing out to ~ 0.02 at $nd \sim 0.528$ (Fig. 3.5B; Supplementary Fig. S3). These network modularity patterns are indicative of more robust age-wise cohesive recruitment in loops than domain structures, with loop prototypes being tightly coupled by age during EF emergence. This trend of course approached an equilibrium as network agglomerative modularity matured and existing forms engaged in widespread recruitment of emerging constructs throughout the timeline. This recruitment trend is also demonstrated in the pairwise NG_{age} heat maps of the EF network by a red sigmoidal signal during early events (first three panels) diffusing into a pervasive (red pixelated) pattern (Fig. 3.5C; Supplementary Video 1). These modular matrix representations along with the insights from the power law and modularity statistical indicators reflect a parallel development of hierarchy that encouraged modules to cluster within modules in a fashion not distinct from the clustered scale-free organization of metabolic networks (47)(Ravasz *et al.*, 2002) (Supplementary Fig. S4; Supplementary Video 2). In contrast, the growth of loop and domain networks while initially driven by recruitment, was quickly but moderately counteracted by age-bound modularity. Thus, the transfer of scale-free properties between loop and domain structures involves a hidden switch to a hierarchically modular structure, which occurred ~ 3.4 Gya. Noticeably, the timing of this switch, as discovered earlier in the literature (14)(Aziz, Caetano-Anollés and Caetano-Anollés, 2016), overlaps with the early development of genetic code specificity in emerging aminoacyl-tRNA synthetases and the ribosome that was moderated by the OB-fold structure (1)(Caetano-Anollés *et al.*, 2013) (Fig.2).

The evolutionary rise of scale-freeness and hierarchical modularity in the emerging EF network of loop prototypes and domain structures is a prediction of the biphasic (bow-tie) theory of module emergence proposed by Mittenenthal *et al.* (2012) to explain concurrent patterns of unification and

diversification existing in biological systems. In the first phase of the biphasic theory, the nodes of the emerging network associate variously, but with weak linkages, through processes of recruitment. As the system grows, nodes diversify by competitive optimization for functional performance. Useful emerging interactions constrain their associations, causing tight linkages to self-organize into modules expressing as closely-knit communities. In a relatively longer second phase, variants of the modules evolve and instigate a new generative cycle of higher-level organization, highlighted by scale-free module recruitment. The network paradigm formalizes the concept of ‘linkage’ by using nodes to represent parts of the system and using links to represent their interaction and/or association. We have previously shown that biphasic patterns are prevalent in dipeptide makeup, loop-mediated flexibility, and size of proteins (Caetano-Anollés et al., 2013; Debès et al., 2013), which is likely linked to their intrinsic disorder. Biphasic patterns were also evident in several biological networks expressing at different time scales, from nanosecond dynamics to billions of years of evolution (Caetano-Anollés et al. 2018). Recently, we uncovered biphasic patterns in evolution of domain organization (Aziz and Caetano-Anollés, 2021). The EF network now showcases its biphasic structuring by integrating communities of interacting structural parts of domains into modular classes of molecular functions. Thus, adaptations to a biphasic pattern of change appear to be a general phenomenon in network biology.

3.3.6 Untangling patterns of molecular innovation and reuse of structures and functions

The waterfall EF, domain and loop network layouts arranged unique time events (228, 226 and 206, respectively) along a timeline that spans from the origin of proteins ($nd = 0$) to the present ($nd = 1$) (Figs. 3 and 4). The distribution and connectivity of loop and domain nodes across these events provides information about how progressive, widespread, and recurrent was the combinatorial recruitment process that embeds loop prototypes into domain scaffolds to generate

new molecular functions. Here, we study loops that are most abundant and widely distributed in genomes, and thus are likely the oldest, following an established rationale (10)(Goncearenco and Berezovsky, 2010). Crisscross patterns in network links strongly suggest recruitment of old loops by younger domains throughout the timeline. In fact, the largest hubs holding most of loop and domain connectivity were observed appearing very early in protein evolution, drawing heavily from innovations appearing during the first 800 million years, but then rigorously extending recruitment from ~2.5 to ~1.25 Gya of protein history (Figs. 3 and 4).

While contemporary co-option of ancient loops and domains was prominent at every time event, most events of recruitment involved loop acceptors originating at $nd=[0.3-0.8]$ and domain acceptors originating at $nd=[0.4-1.0]$ (Fig. 3.4). These patterns were made quantitative by dissecting the source-sink relationships and evolutionary span of network connectivity with bar plots describing the chronological cumulative connectivity along the timeline (Fig. 3.6). The networks showcase an overwhelming majority of modern recruitment events, some very recent, with relatively younger sink nodes ($nd=(0.5-1.0]$) being acceptors of very old donors or source nodes originating at $nd < 0.3$. In this respect, sink loops seem to be particularly adaptive, progressively drawing innovation from donors spanning the entire timeline. Box-and-whisker plots of cumulative weighted indegree and outdegree across network chronology (Supplementary Fig. S5) and scatter plots with linear distribution models of degree totals at $nd=1$ (Supplementary Fig. S6) provided further insight into the patterns of contraction and expansion of mutually adaptive loop and domain innovations. Specifically, individual domains took advantage of the repertoire of very ancient donors for their functional tasks and showed signs of co-option among modern domains late in evolution ($nd \geq 0.8$), confirming evolutionary patterns of recruitment obtained in the enzymatic analysis of metabolic networks (58)(Kim et al., 2013). Similar patterns

were identified when exploring the evolution of protein domain organization (59)(Aziz and Caetano-Anollés, 2021) and EFL-mediated elementary functions (14)(Aziz, Caetano-Anollés and Caetano-Anollés, 2016), where most innovation happened during the first ~1.8 Gy of protein history (Fig. 3.6). Analysis of the highly connected loop and domain subnetworks of the EF network projections showed that although the largest hubs appeared early in evolution, the recruitment process was generative of new hubs throughout the modern structural world (Fig. 3.7; Supplementary Videos 3 and 4).

Finally, the connectivity of loop and domain components of the EF network gradually evolved from 1 to a global average of 1.77 (± 0.032) loops per domain and 1.31 (± 0.018) domains per loop, respectively (Supplementary Fig. S7). Remarkably, loop connectivity fluctuated up (~1.7 domains per loop), down (~1.2) and up again (~1.3) during early protein evolution ($nd < 0.2$). Domain connectivity fluctuated in a mirror fashion but with a slight phase delay ($\Delta nd = 0.02$) and with peaks of ~1.8, ~1.5 and ~1.7 loops per domain. These dual hourglass trends suggest some frustrated dynamics of growth, in which early loop co-options were more vigorous than trends towards economy in the number of loops in typical active sites of domain structures.

The connectivity patterns we identified expose a fluid emergence of domain structures and their associated molecular functions, falsifying some alternative hypotheses. Historically, the progression of molecular innovation most likely involved ligation of dipeptides or small polypeptides with limited ordered structure, followed by the formation of larger peptides harboring stable loops (1,31)(Caetano-Anollés et al., 2012, 2013), and finally the combination of loops to form defined folded topologies typical of small protein domains (7,8,10)(Berezovsky and Trifonov, 2001; Aharonovsky and Trifonov, 2005; Goncarenko and Berezovsky, 2010). While a continuum of these prior forms is expected according to the principle of spatiotemporal continuity,

our analysis revealed that the adaptive formation of useful loops and domains occurs at different rates in ripples and waves of recruitment and innovation. The early appearance of functional loops provides raw materials for loop combinatorics in the formation of domains along the entire timeline of protein history. Nevertheless, introduction of novel loops and domains seems ongoing and their involvement in recent combination events with older domains suggests that old loops are still evolutionarily active and evolvable forms instead of relics headed for extinction. Thus, the quick establishment of highly conserved sequence motifs in elbow regions of loops and their combinatorics defined a hierarchical structural complexity, tightly interrelated throughout protein history. Remarkably, phylogenetic studies demonstrate similar dynamics materializing with domain structures and their combinatorial use in multidomain proteins (22,59)(Wang and Caetano-Anollés, 2009; Aziz and Caetano-Anollés, 2021). These connectivity patterns notably falsify alternative evolutionary scenarios of sequential and separate build-up of loop and domain repertoires. Loops and domains seem to have evolved in concert, but at different rates. Although highly conserved evolutionarily, our results falsify a ‘molecularly canalized’ model of these forms that resist evolutionary change. Specifically, the loop prototypes continue to arise from diverse families of sequence motifs, expressing themselves in different structural contexts. Historically, only functionally useful prior forms prevailed when they provided properties that extend the persistence of the emerging cells, including membrane stability, transport modulation, bioenergetics, and peptide-cofactor biosynthetic functions.

3.3.7 Two primordial recruitment waves

Two primordial waves of functional innovation arising from ancient ‘*p*-loop’ and ‘winged helix’ domains were originally identified in metabolic pathways (60)(Caetano-Anollés et al., 2007). Later evolutionary studies of elementary functionome (14)(Aziz, Caetano-Anollés and Caetano-Anollés,

2016) and protein domain organization (59)(Aziz and Caetano-Anollés, 2021) also uncovered these same two waves of innovation. Remarkably, the waterfall diagrams of the modular EF network and its projections (Figs. 3 and 4) revealed that these ancient recruitment pathways arising from ‘*p*-loop’ and ‘winged helix’ domains were also presents in our networks, uncovering separate origins of sandwich, barrel and bundle domain structures. The versatility of the waterfall visualization in the form of highly connected (reduced) subnetworks of the hubs in the loop and domain projections visually untangled the two recruitment waves (Fig. 3.7). The realization that these evolutionary patterns are parallelly uncovered with various data sources, as depicted by simulated movies as well (Supplementary Videos 5, 6 and 7), is remarkable, and strongly supports the historical statements we here elaborate.

The first larger wave originated in the *p*-loop containing nucleoside triphosphate (NTP) hydrolase domains (c.37.1.12, c.37.1.8 and c.37.1.1) (Figs. 4A and 7A) and their contemporary and relatively long *p*-loop-related DS.EH.6.1.1 and DS.EH.6.1.4 prototypes, with eight relatively recent terminal loop prototypes, DS.EH.0.1.17, DS.EH.6.1.2, DS.HE.3.70.1, DS.EH.2.17.1; DS.BN.2.3.4; DS.GH.2.2.2, DS.HE.0.1.1 and DS.GH.2.1.4, respectively (Figs. 4B and 7B). These domain families of the P-loop containing nucleoside triphosphate hydrolase (c.37.1) superfamily are the most ancient and most popular Rossmannoid $\alpha/\beta/\alpha$ -layered domain structures of the timeline of domain history (19,20,31,43)(Caetano-Anollés and Caetano-Anollés 2003; Caetano-Anollés et al., 2007, 2009, 2012). The *p*-loop prototype was crucial in enabling the nucleotide triphosphate binding functions of the *p*-loop hydrolase fold with its Walker A (*p*-loop) sequence motif located at the elbow connecting the first β -strand of the main β -hairpin loop (or loop derivatives), which binds to di- and tri-nucleotides. The EF network confirmed that the ‘*p*-loop’ wave instigated many massive pathways of loop recruitment over ~2.5 billions of years (Gy) of history, involving mostly

cysteine-rich loop prototypes. In these strong recruitment pathways, the most ancient domains such as NAD(P)-binding Rossmann-fold domain (c.2.1.2) family and the S-adenosyl-L-methionine-dependent methyltransferase domain (c.66.1.43) family with their 3-layered $\alpha/\beta/\alpha$ structures, and the ancient OB-fold of the nucleic acid-binding protein domains (b.40.4.5 and b.40.4.4) with their closed or partly-opened β -barrel structure, enabled a plethora of functions related to metabolism and translation. In particular, the cysteine-rich metal binding loop of the secondarily connected, downstream DS.HE.3.1.1 prototype formed a cysteine nest that coordinated Zn^{2+} metal binding necessary for interactions with nucleic acids in 6 loop-related domains. Among other notable domains of this wave were the class II aminoacyl-tRNA synthetases and biotin synthetases (d.104.1.1) and nucleotidyltransferase (d.218.1.5) families with $\alpha/\beta/\alpha$ -layered and sheet structures, and more derived beta and beta-prime subunits of DNA dependent RNA polymerase (e.29.1.1 and e.29.1.2), and prokaryotic type I DNA topoisomerase (e.10.1.1) families with β -barrel and winged helix-like structures (Fig. 3.4A). The terminal DS.HE.2.1 prototype group of the cysteine-rich loop recruitment pathway completed the tRNA-independent cysteine biosynthetic pathway 3-3.2 Gya by providing functions to the tryptophan synthase β -subunit-like PLP-dependent domain (c.79.1.1) of serine acetyl-transferase and O-acetylserine sulfhydrylase enzymes (58)(Kim et al., 2012). This probably enhanced cysteine availability for cysteine-rich loop recruitment and binding of Fe-S clusters, which started with the PLP-dependent transferase c.67.1.1 domain 3.5 Gy-ago. Finally, the DS.BN.5.5.1 prototype hub also linked downstream glycine and glutamate-rich DS.BK.4.63.1 and DS.BN.6.15.1 prototypes and the upstream glycine-rich nucleotide-phosphate binding DS.HH.1.1.1 that is typically embedded in β/α -barrel structures widespread in metabolism via the Rossmann-like tyrosine-dependent oxidoreductases (c.2.1.2) family structure. Loops DS.HH.1.1.1

acted as hub for other downstream prototypes, including DS.EH.4.2.1 and DS.HH.1.1.3 (Fig. 3.4B).

The second wave in turn originated in the ‘winged helix’ DNA-binding domain (a.4.5) superfamily (Figs. 4A and 7A), which resurfaced throughout the timeline, starting with the MarR-like transcriptional regulators (a.4.5.28) family ~3 Gya and its contemporaneous loop hub, the DS.HH.3.1.1 prototype (Figs. 4B and 7B). The wave appeared soon after the *p*-loop wave but was somewhat constrained in scope; part of it merged with the *p*-loop wave through the sister loop hub, the DS.GH.0.0.1 prototype, and its ancestral hub domain, the N-acetyl transferase (d.108.1.1) family. The a.4.5 superfamily harbors the DNA/RNA-binding 3-helical bundle fold (a.4) structure, which is flanked by a 4-strand β -sheet. This domain exposes crucial elbows structures between the helix-turn-helix motifs, harboring the specificity of protein-protein and protein-RNA interactions typical of these enzymes. The winged domain has a central role in transcription, providing flexibility and nucleic acid clamping capacity to RNA polymerases (61)(Teichmann, Dumay-Odelot and Fribourg, 2012). These domains also provide surfaces for paired structural recognition in complexes (like polymerases, ubiquitin-ligases, condensins) and other domain interactions.

Remarkably, these two waves of the EF network and its projections denote the same primordial sandwich $\alpha/\beta/\alpha$ -layered structures, β -barrels and helical bundle structures referred earlier as part of the first 54 historical domains in evolution (31)(Caetano-Anollés et al., 2012). Also, the ‘*p*-loop’ and ‘winged helix’ waves notably embed the first two major gateways of enzymatic recruitment we identified earlier in metabolism, the first gateway mediated by the c.37 fold that originated in the energy interconversion pathways of the purine metabolism subnetwork, and the second mediated by the a.4 fold that originated in the subnetwork of porphyrin and chlorophyll metabolism, and the biosynthesis of cofactors (32,60,62)(Caetano-Anollés et al., 2007; Caetano-

Anollés et al., 2009; Caetano-Anollés & Caetano-Anollés, 2013). Obtaining these concordant results using different structural and evolutionary data sets supports the historical statements we here propose.

3.3.8 Modeling the origin and evolution of the most ancient domain structure

Phylogenomic data-driven chronologies and networks describe how evolution embeds loops into protein domains. Remarkably, this information allows to model the emergence of folded domain structure by first determining the sequence of events of loop recruitment and then using deep learning algorithms of *ab initio* structural prediction to find evolutionary patterns of convergence towards the central structural core of the folds.

While this is still a work in progress, we were able to illustrate the power of this approach by focusing on the most ancient domain family, the ATP-binding cassette (ABC) transporter ATPase domain-like (c.37.1.12), which is part of the P-loop containing nucleoside triphosphate hydrolases fold (c.37) and superfamily (c.37.1) of SCOP. The fold has a 3-layered $\alpha/\beta/\alpha$ sandwich arrangement with parallel or mixed β -sheets of variable sizes and topologies. The P-loop containing ABC transporter family that is responsible for the transport of a wide range of molecules across membranes (from small compounds to polypeptides) has a central core with a RecA topology that is missing some typical secondary structures of the fold. Modeling the birth of the fold demanded three procedural steps. First, the times of origin of loops were traced onto the 3-dimensional structure of a representative ABC transporter molecule, as we have previously done with proteins (31)(Caetano-Anollés et al. 2012), protein complexes (X)(Caetano-Anollés and Seufferheld, 2013), or the ribosome (X)(Harish and Caetano-Anollés, 2012). Second, a time-ordered series of growing molecules was reconstructed by stitching loop sequences together,

starting with the most primordial loop (the P-loop) and adding loops sequentially according to their time of origin. Finally, the three dimensional structures of the growing molecules were modeled directly from their sequences with AlphaFold2 (X)(Jumper et al., 2021), the star of the last biannual structure prediction experiment (CASP, round XIV) (X)(Kryshtafovych et al., 2021). AlphaFold2 uses a deep learning algorithm to predict 3-dimensional structure directly from its sequence with levels of accuracy that are within the margin of error of experimental structure determination methods. Calculation of the median ‘global distance test’ (GDT), which measures the similarity of predicted and experimentally acquired structures with known amino acid correspondences, resulted in total scores of well above 90%, indicating global folds and structural details were correct. AlphaFold2 extracts co-evolutionary information in both multiple sequence alignments and structural templates from libraries using an ‘oracle’ that can quickly and iteratively identify which alignment and ‘pair representation’ of structural template data is more informative. This neural network-generated information is then processed by an ‘Evo former’ module to produce increasingly refined deep learning models of both sequence and structure, which converge into the structural prediction. The module uses two attention matrix-based ‘transformer’ architectures to convert the discrete vocabulary of sequence alignments into a continuous ‘embedded’ space of structure capable of training the multiple-layered neural networks. The structural prediction is finally assembled by a ‘structure’ module, which considers a protein as a ‘residue gas’. Each amino acid is modelled as a floating triangle with the three atoms of the backbone, which coalesce into the structure by translations and rotations in space using another attention transformer mechanism and ulterior refinements.

Fig. 8 illustrates the results of the three-step strategy. The loop prototypes of the P-loop ATP-binding domain were traced onto the crystallographic atomic structure of a histidine permease

enzyme by coloring loop substructures according to the times of origin of their corresponding prototypes (Fig. 3.8A). These tracings represent a model of accretion of loop substructures in the permease molecule, which in itself becomes a model of structural evolution. The timeline of accretion began with the oldest loop of the molecule, the P-loop (loop structure 34), which mapped to the DS.EH.6.1.1 prototype. The evolutionary growth of the protein was represented as a series of insertions of loop structures in the form of a series of loops [34>213>80>186>....] or prototypes [DS.EH.6.1.1>DS.HE.2.2.4>DS.EH.2.1.58>DS.HE.4.2.20>]. Alternatively, molecular growth was more appropriately described as a series of molecular intermediates [34 (nd = 0), 34|213 (nd = 0.112), 23|80|213 (nd = 0.146), 24|80|186|213 (nd = 0.184), ...], with loop adjacencies in sequences represented with pipe symbols and age (nd) of intermediates given in parentheses (Fig. 3.8B). This loop sequence representation of growing molecules allows to both track locations of loop insertions at every time step and translate a loop sequence into a series of sequences for AlphaFold2 input. Finally, the *ab initio* structural predictions produced a series of high-resolution structural representations of the growing molecules, which were placed within a geological time scale framework (Fig. 3.8C). The per-residue confidence estimates of AlphaFold predictions, ranged 62.3-95.0, with values increasing with protein length, show confident to very high confident predictions. Since pLDDT is particularly useful for assessing local structural accuracy or disorder, we plan to show confidence variation along the protein chain (useful for identifying highly flexible or disordered regions), structural alignments of the 5 ranked predictions produced by the software, and predicted alignment error (PAE) plots measuring confidence in the relative positions of pairs of residues, which is important when assessing domain packing and large-scale topology.

Remarkably, the series of predicted structural intermediates converged towards a mixed β -sheet flanked by α -helices making up the ‘binding’ cassette of the primordial RecA-like domain core (shaded region of the timeline, Fig. 3.8C). This occurred within a period of 700 million years spanning 3.8-3.1 Gya (nd =0-0.184). A three-stranded antiparallel β -sheet already appeared 3.2 Gya in the 3-loop intermediate (23|80|213) hosting the DS.EH.6.1.1, DS.HE.2.2.4 and DS.EH.2.1.58 prototypes, but its structure was likely fluid given analysis of residue pairs in PAE plots and pLDDT variation along the chain (planned). The lone α -helix of the structure belonged to the bracing secondary structures of the P-loop. Further addition of the DS.HE.4.2.20 prototype 3.1. Gya (nd=0.184) rearranged the molecule by adding two α -helices to produce a sandwich structure but converting the initial antiparallel arrangement into a stable parallel β -sheet topology. As accretion proceeded, the fold continued to be accessorized with β -strands and α -helices, which solidified the typical 3-layered $\alpha/\beta/\alpha$ fold. An extra terminal β -strand in antiparallel arrangement was added 2.7 Gya (nd=0.245) by incorporation of the DS.BN.2.1.7 prototype, while the previously incorporated DS.HE.2.2.4 prototype gained its C-terminal β -strand ultimately producing a 5-strand mixed β -sheet. PAE plots showed that the 23|80|186|213|222 molecular intermediate appearing 2.7 Gya exhibited a cohesive domain structure. The extra 213 loop structure eliminated the two error-prone bands present in the PAE plots of the prior molecular intermediate (planned). In a next step, an additional β -strand and α -helix were added to the molecule 2.5 Gya (nd=0.322) following the integration of the first DS.GH.2.2.2 prototype of the helical bundle. This crucial step completed the 6-strand central β -sheet of the extant molecule.

During the initial structural convergence process, there was a temporal sequence of bracing structures of the loops that obeyed molecular elongations matching secondary structures already

in place. The sequence followed EH>HE>EH>HE>BN, only stopping by the evolutionary appearance of the first helical component of the bundle appeared 2.5 Gya (nd=0.322). This steady pattern of ‘reformation’ follows a cryptic phenomenon illustrated by the appearance of the loop structure 213 (mapping to DS.HE.2.2.4) as a helix-coil region 3.3 Gya (nd=0.122). This integrated structure was reformed into a fluid beta-hairpin 3.2 Gya (nd=0.146) when pushed towards the C-terminus by the insertion of loop 80 (DS.EH.2.1.58). Its integration onto the expanding β -sheet however was only stabilized into its final form HE, 3.1 Gya (nd=0.184), once the insertion of loop 186 (DS.HE.4.2.20) reformed the terminal loop 213 structure placing the terminal β -strand at the C-terminal region of the β -sheet. A similar phenomenon occurred 2.4 Gya (nd=0.373). following the integration of the first α -helix of the bundle. The incorporation of loop 68 (which mapped to DS.BK.2.1.3) downstream the P-loop structure resulted in the formation of a loop form HE, mimicking the DS.HE.0.1.1 prototype (colored green) that was integrated 400 million years later (2 Gya; nd=0.472). This instance of loop reformation from a BK to HE bracing architecture seems to represent an instance of significant structural rearrangement. These types of rearrangement continue throughout the timeline but are particularly striking during the past 50 million years (nd=0-0.991) when an entire 4-strand β -sheet was formed.

Convergence towards the formation of the ‘binding’ cassette of the primordial ABC transporter required a P-loop-centered nucleation of only three loop structures (213, 80 and 186), which are relatively far away from each other in the extant sequence and structure. To test if convergence was resilient, we conducted reshuffling experiments based on the 34|80|186|213 loop sequence where we systematically replaced loop 80 in the second position by all possible loops (60, 68, 72, 85, 110, 126 and 151) that would maintain sequence order, and separately loop 213 in the fourth position by the only option, loop 222 (note that loop 186 could not be replaced). We then modeled

structures from the reshuffled sequences and compared them to the reference structure using pruned and total RMSD measurements of structural overlaps (planned). In all cases, reshuffling increased RMSD values significantly, despite changing only one loop in the set of 4 in the experiment. Note that reshuffling with loops 68, 72 and 151 (with ages 1.2-2.5 Gya) destroyed completely the β -sheet configuration, while 60, 85, 110 and 126 (with ages 0.05-2.0) preserved it. Reshuffling of two loops in the set of 4 destroyed completely the core structure (planned). These experiments show that the phylogenomic-informed temporal sequence A_34|80|186|213 is very sensitive to loop composition, falsifying the notion that a converge towards a fold can occur at random. The convergence phenomenon is thus real and not an artifact.

3.4 Conclusions

The library of single domain protein structures is essentially complete (Skolnick et al., 2012), and so is the library of loop prototypes, which has all geometries sampled (X)(Fernandez-Fuentes et al. 2010) and is considered saturated even for the case of long loops (X)(Bonet et al., 2014bis). These properties guarantee both an exhaustive phylogenomic exploration of the history of domains and loops but also deep learning's ability to map sequences to structures and solve the fold recognition problem in *ab initio* explorations. Here, we trace the history of growing EF networks and their projections, verifying the existence of two primordial waves of functional innovation in elementary functional loops involving founder 'p-loop' and 'winged helix' domain structures (14)(Aziz, Caetano-Anollés and Caetano-Anollés, 2016). These waves originally explained recruitment patterns responsible for the origin of modern metabolism (60)(Caetano-Anollés et al., 2007). Remarkably, we find that the dynamics of recruitment of loop prototypes into domain structures is ongoing and highly modular, but that there is also an underground recruitment process

of non-modular loop structures that are drawn at each time event of the timeline. The evolving EF network became structured in evolution by exhibiting hierarchy, modularity, and a power law-based underlying scale free behavior. Integrated communities of interacting structural parts of domains defined modular classes of molecular functions in biphasic patterns of emergence. Collectively, links encapsulated the growth of an elementary alphabet of functions in loops embedded in an alphabet of structures in domains.

To explore the interplay of these two molecular languages, we modelled the evolutionary emergence of folded structure using historical information derived from chronologies and networks and deep learning algorithms of *ab initio* structural prediction. We focused on the oldest domain of the timeline, the P-loop domain of ABC transporters. Remarkably, as accretion proceeded in evolution, the fold converged relatively quickly towards its typical core structure. Convergence, which was organized around the nucleotide-binding P-loop prototype, first materialized into a 3-strand β -sheet, then into the 3-layered $\alpha/\beta/\alpha$ structure, and finally into an extended central β -sheet forming first a 5-strand and then a 6-strand planar structure. Remarkably, a recent phylogenetic-inspired engineering exploration was able to generate small P-loop-containing loop proteins capable of binding a range of phosphate-containing ligands, including RNA and single stranded DNA (X)(Romero et al., 2018). The P-loop prototypes were embedded in a fold made of four tandem β - α repeats with a 3-layered $\alpha/\beta/\alpha$ sandwich architecture, which much resembled the structural intermediates that appeared 2.7-3.1 Gya in our evolutionary timeline. The study confirmed that short (55-residue) P-loop proteins were catalytically active, supporting the functionality of the early molecular intermediates.

Structural convergence suggests a primordial folding vocabulary exists in loop structures that overrides the stochastic effects of recruitment. This vocabulary is likely driven by structural reformations occurring within a combinatorial (syntactic) landscape of innovation. Our analysis prompts generalizing *ab initio* structural prediction of molecular intermediates to all domains in an exploration of how semantics (the meaning of functions and structures) determine the pragmatics (context-dependent rules) of molecular communication.

3.5 Methods

3.5.1 Phylogenomic analysis

We conducted a census of structural domains defined at domain level of SCOP classification (Murzin et al., 1995) in 8,127 proteomes of completely sequenced genomes from cellular organisms and viruses using profile hidden Markov models (Eddy, 2011) of the SUPERFAMILY database with assignments at $e < 0.0001$ (Gough et al., 2001). Proteomes analyzed were from 139 archaeal, 1,734 bacterial, and 210 eukaryal representative and reference cellular organisms sampling all major taxonomic groups of the RefSeq database (O’Leary et al., 2016). We also included a set of non-redundant 6,044 viral proteomes after exclusion of unclassified viruses, which were categorized according to genome type and viral replication using the Baltimore classification. Phylogenomic analysis was carried out as previously described (Nasir and Caetano-Anollés, 2015) according to verified phylogenomic protocols (Kim et al., 2012; Kim & Caetano-Anollés, 2012). Briefly, the abundances of individual domains in every proteome were log-transformed and rescaled by dividing by the maximum abundance in that proteome followed by additional rescaling to 24-character states to ensure compatibility with PAUP* (version 4.0) (Swofford, 2022). Normalization and rescaling protected against the effects of unequal proteome

sizes and variances. Maximum parsimony (MP) was used to reconstruct optimal phylogenomic trees of domains describing the evolution of domains (taxa) using proteomes as characters. Trees were rooted using the Lundberg method in PAUP*, which does not require specification of any outgroup taxon and complies with Weston's generality criterion of phylogenetic rooting. Instead, an unrooted network is calculated, which is rooted a posteriori by the branch yielding a minimum increase in tree length. We note that MP approximates maximum likelihood when phylogenetic characters evolve at different rates (Kolaczowski and Thornton, 2004) and is appropriate for global proteome studies. Bootstrap analysis with 1,000 replicates was performed to assess the reliability of deep evolutionary relationships. Trees were visualized with FigTree version 1.4.4 (available at <http://tree.bio.ed.ac.uk/software/figtree/>). A single tree of domains reconstructing the evolutionary history of 3,892 domains in the 8,127 proteomes, which was recovered from a heuristic search (Fig. 3.1C), was used to determine the times of origin of domains by calculating *node distance* (*nd*) values directly from the rooted tree. A calibrated molecular clock of domain structures was used to calculate geological age in Gy (Wang et al., 2011).

3.5.2 Time of origin assignments

Since a loop is embedded in a domain structure and both loops and domains describe functional and structural abstractions, the age of domains can be directly transferred to loop prototypes. Whenever an older domain donates its loop to a younger domain in evolutionary recruitment, the loop can neither be younger than the younger domain nor be older than the older domain. Consequently, we considered two likely schemes of age transfer: (i) the age of a loop is the age of the most ancient associated domain, or (ii) the age of a loop is the age of the more recent of the pair of most ancient associated domains. These schemes consider the age of a loop prototype as either the age of the first structural scaffold or the age when the loop function is first transferred

between structural scaffolds, respectively. Both schemes provided similar age mappings. For that reason, we only present mappings derived using the second more conservative scheme. Since the first loop that appeared in evolution must generate the first domain in order to preserve the '*lex continui*' principle and a donor loop has to be either older or at least contemporaneous to the acceptor domain, the most ancient loop DS.EH.6.1.1 was assigned a time of origin of 0 according to scheme (i) as an exception because of the absence of any older donor domain.

3.5.3 Domain and loop prototype data

Loop prototypes were computationally identified by filtering DS-derived loop structures from ArchDB (Bonet *et al.*, 2014) while mapping domains to loops at e-value <0.001. This resulted in 88,321 loops structures clustering into 7,078 unique loop prototypes that mapped to 2,447 domains. Note that each loop structure in ArchDB has one loop prototype annotation in the DS classification system with many-to-many mappings between loops and domains. Out of the set of 7,078 loop prototypes, a subset of 5,125 only mapped 'horizontally' and uniquely to domains of a same age within a set of 1,965 domains. They were reported as '*non-modular*' (NM) loop prototypes because they failed to be recruited by different domains across the timeline, including those that were contemporaneous. In contrast, a subset of 1,937 loop prototypes mapped 'vertically' to 1,442 domains with times of origin spread throughout the timeline. They were reported as '*modular*' (M) loops since they acted as modular units of structural, functional, and evolutionary significance. Finally, a subset of 16 loops mapped 'horizontally' to sets of 2 contemporaneous domains, 32 in number. We reported these loops as '*modular*' (contemporaneous) (M') loops because they involved recruitments occurring within individual time events and representing focal innovations.

3.5.4 Network visualization and analysis

Networks were visualized and analyzed using Pajek (Mrvar and Batagelj, 2016) and R's *igraph* package (Csardi and Nepusz, 2006). Community-based layouts of the networks were generated using the Visualization of Similarity (VOS) clustering method (Van Eck and Waltman, 2007; Waltman, van Eck and Noyons, 2010). Network properties were analyzed with code constructed using graphing packages and tools of R (Ihaka and Gentleman, 1996; R Core Team, 2014). A detailed description of data files, partitions and functions used to analyze network data, produce charts and plots, compute power law statistics and modularity indices, and construct waterfall diagrams can be found in the Supplementary Materials and Methods.

3.5.5 Statistical Analysis

3.5.5.1 Power law network behavior

Scale free network behavior was studied using $P(k)$ vs. k (probability of having k -neighbors vs. k) and $\log\text{-}\log$ (log of $P(k)$ vs. log of k) mathematical curves, with linear regression models to derive γ of the power law and the determination coefficient (R^2). γ is the absolute slope of the log linear model. Higher γ indicates increased tendency towards preferential attachment. R^2 describes the percentage of the data fitting the linear model. When both γ and R^2 are high, scale free behavior should be considered strongly supported. Other power law statistics included: (i) KS fit statistic, which compares the fitted distribution with the input degree vector; (ii) the KS p -value, with the null hypothesis of data being drawn from the power law distribution (Clauset et al., 2009; Newman, 2005); and (iii) the exponent of the fitted power law distribution (α), which assumes $P(X=x)$ is proportional to $x^{-\alpha}$. Lower KS fit score, larger KS p -value (≥ 0.05), and higher α suggest better fit to power law distribution. We also determined the maximum log likelihood of the fitted

parameters. Reference networks were created using ‘Barabási’ (Barabási and Albert, 1999) methods of R’s *igraph* package (Csardi and Nepusz, 2006) to simulate power law and extended age-dependent control models for the corresponding networks.

3.5.5.2 Network modularity

We studied modularity with six indices: (i) The VOS Quality index (VQ), was generated by the Pajek layout algorithm that considers weights of links (edges/arcs) as similarities. Communities were iteratively drawn closer based on similarity and the quality index of the final layout with least crossings and closest clusters was given. VQ is then calculated as $\sum_{i=1}^c \sum_{j=i+1}^c (e_{ij} - a_i^2)$, where c is the number of communities. e_{ij} is the fraction of edges with one node v in community i and the other w in community j , given as $\sum_{vw} (A_{vw}/2m)$ with $1_v \in c_i$, $1_w \in c_j$, where m is the sum of weights in the graph and A_{vw} = the weighted value or 0, indicating presence or absence of edge between the nodes v and w , respectively, in the adjacency matrix A of the network. Finally, a_i is the fraction of weighted k neighbors that are attached to the nodes of a community i , i.e $k_i/2m$ (Van Eck and Waltman, 2007; Waltman, van Eck and Noyons, 2010); (ii) The Clustering Ratio (C-ratio) is the ratio of the number of clusters to the size of an inter connected node set; (iii) The average Clustering Coefficient (C) describes the mean ratio of triangles to connected triads over all nodes in the simplified (undirected/unweighted) network (Wasserman and Faust, 1994; Ravasz *et al.*, 2002; Barrat *et al.*, 2004). C is meaningful only for unimodal graphs (Newman, Strogatz and Watts, 2001). We also report coefficients of linear regression over C for domain and loop network projections; (iv) The Fast Greedy Community (FGC) hierarchical agglomeration algorithm detects community structure with linear run time $O(m d \log n) \sim O(n \log^2 n)$, of a network with m edges, n nodes, and depth d of the dendrogram describing its community structure (Clauset, Newman and Moore, 2004); and (v and vi) The *Newman-Girvan* algorithm index (NG), computed with partitions

defined by age (NG_{age}) and VOS clustering (NG_{vos}). NG calculates the modularity of a network based on some classification (partition) to measure how good that classification is in dividing the various node types, indicated by assortative (positive) or disassortative (negative) mixing across modules (Newman and Girvan, 2004). NG equals $1/(2m)\sum_{ij}(A_{ij}-1/(2m)k_i k_j*\Delta(c_i,c_j))$, where m is the collective weights in the graph, A_{ij} are weighted entries in the adjacency matrix, k_i , k_j and c_i , c_j are the weighted degrees and the components (numeric partitions), respectively, of nodes i and j each, and finally, $\Delta(x,y)$ is 1 if $x=y$ and 0 otherwise (Newman and Girvan, 2004). VQ , C -ratio, C and FGC each range from 0 to 1, while the NG indices range from -1 to 1. Higher indices represent strong network modularity at an event. Heatmaps were generated using customized scaled modularity matrices with elements given as $(A_{ij}-k_i k_j/(2m))M_{nd}$, where A_{ij} , k_i , k_j and m are as defined for NG (Newman and Girvan, 2004), and M_{nd} is a network's modularity index at event nd . Dendrograms were calculated as squared Euclidean distance matrices indicating dissimilarities between the cluster means (Borg and Groenen, 2003). The distance (or dissimilarity) matrices were hierarchically clustered with the Ward's minimum variance method aiming at finding compact, spherical clusters (Murtagh and Legendre, 2014).

3.5.6 *Ab initio* modeling

The 3-dimensional structures of evolving molecules were modeled directly from their sequences with the AlphaFold2 pipeline (X)(Jumper et al., 2021). We requested output of 5 ranked models obtained with 3 recycles using PDB70 as template and the multiple sequence alignment (MSA) mode MMseqs2 (UniRef100+Environmental). The use of PDB70 template did not significantly affect modeling results. Accuracy was measured with pLDDT and the predicted aligned error (PAE). pLDDT provides a per-residue estimate of prediction confidence based on the local Distance Difference Test (lDDT)- $C\alpha$ metric (X)(Mariani et al., 2013). The expected prediction

reliability of a given region or molecule follows pLDDT ‘confidence bands’: >90, models with very high confidence; 90-70, models with confidence, showing good backbone predictions; 70-50; models with low confidence; and <50, models with very low confidence, generally showing ribbon-like structures. pLDDT < 60 can be considered a reasonably strong predictor of intrinsic disorder. TAE measures confidence in the relative positions of pairs of residues and is a good metric to evaluate the cohesiveness of domains. structural alignment subroutines of the visualization software Chimera (X) (Pettersen et al., 2004; available at <https://www.rbvi.ucsf.edu/chimera>).

3.6 Figures

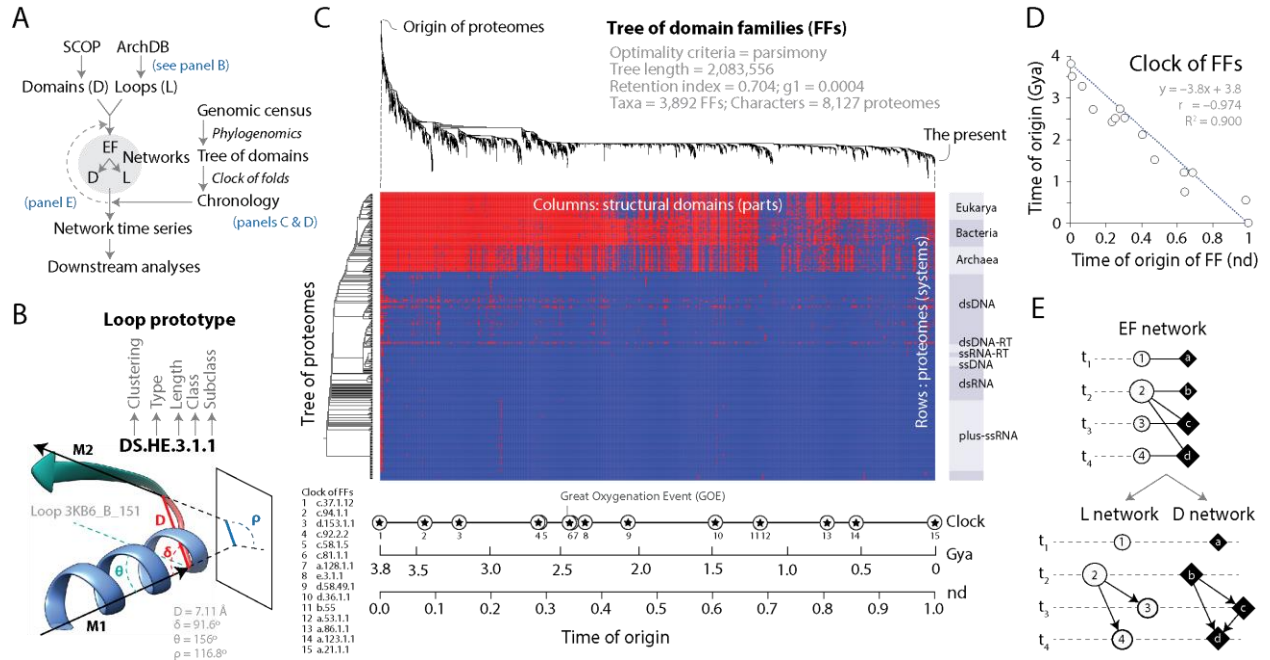


Figure 3.1. General experimental strategy. (A) Workflow describing the generation of time series of a bipartite ‘elementary functionome’ (EF) network and its loop (L) and domain (D) projections. The SCOP and ArchDB classifications are used to map loop prototypes to domain families along a chronology built with phylogenomic methodologies. The chronology adds time to network makeup and downstream analysis evaluates network structure (e.g., hierarchy, community structure) and fold emergence with AlphaFold predictions. (B) Definition of a loop prototype in ArchDB. The loop is defined by the bracing secondary structures of the loop, the number of residues forming the aperiodic structure, its conformation (ϕ and ψ backbone dihedral angles of the participating residues), and the geometry of the loop. The atomic model of the 3KB6_B_151 loop that is part of prototype DS.HE.3.1.1 shows its geometric properties defined by four internal coordinates (D, δ, θ, ρ) extracted from the orientation of principal vectors (M1 and M2) of bracing secondary structures: D (Distance), the Euclidean distance between the boundaries of the aperiodic structure; Δ (hoist) angle (δ), the angle between M1 and D ; Θ (packing) angle (θ), the angle between M1 and M2; and ρ (meridian) angle (ρ), the angle between M2 and the plane Γ defined by the vector M1 and the normal to the plane formed by M1 and D . (C) Phylogenomic tree of structural domains reconstructing the evolutionary history of 3,892 fold families (domains) in 8,127 proteomes. The evolutionary heat map describes the phylogenetic data matrix of genomic abundances used to build the tree, with domains ordered according to their time of origin in a relative scale (nd) and rows describing the 8,127-proteome set ordered according to a rooted tree of proteomes. The domain abundance is described with a scale. (D) A molecular clock of folds established that nd values of domains were linearly correlated with geological time in billions of years (Gy) using Pearson ($r = -0.974$, $p < 0.00001$) and Spearman ($\rho = -0.961$, $p(2\text{-tailed}) = 0$). $\rho < r$ rejects non-linear behavior. (E) Diagram illustrating an undirected bipartite EF network and its L and D projections unfolding along time events (t_1, t_2, t_3 and t_4). Nodes are described with symbols (circles = loops, rhomboids = domains), with size proportional to the number of links they establish.

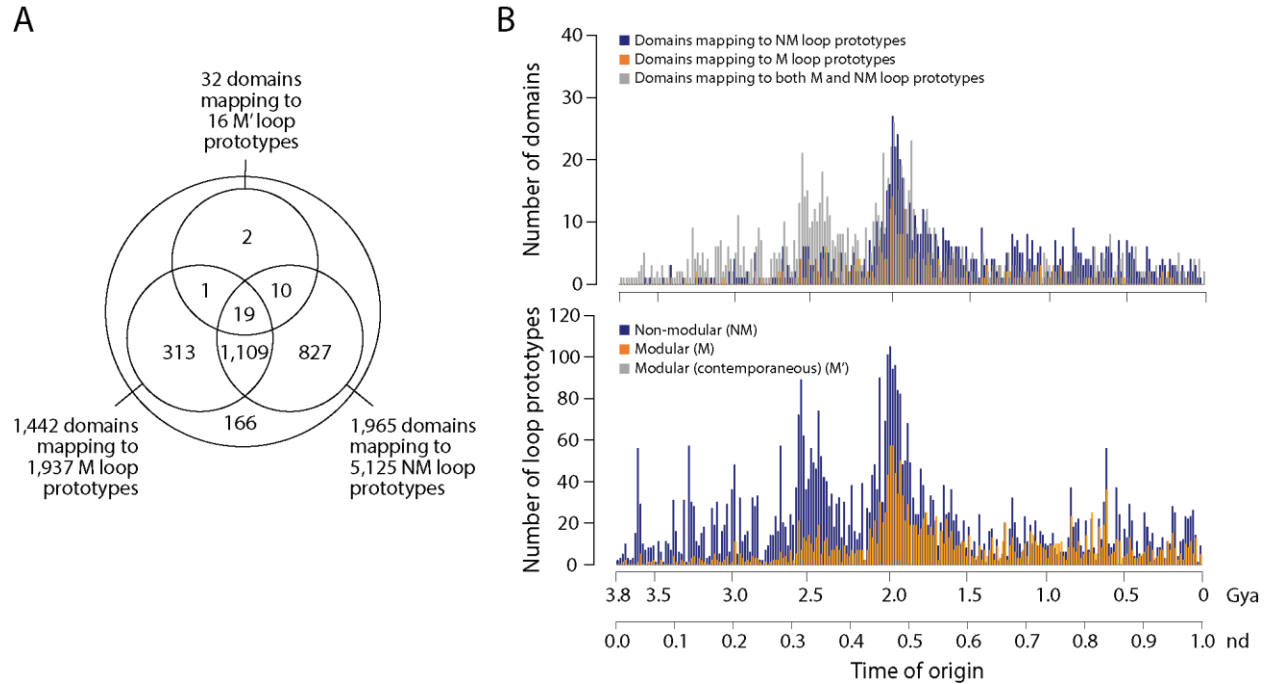


Figure 3.2. Mapping loop prototypes to domains along the domain chronology. (A) Venn diagram describing how structural domains map to modular (M and M') and non-modular (NM) loop prototypes. The set of 166 domains does not map to loops but can be traced to a set of 228 filtered M loops. (B) Tracing domains and loops along the evolutionary timeline. The gradual appearance of domains belonging to the three most populated groups of the Venn diagram along domain history reveals the early rise of domains mapping to both M and NM loops in evolution. The gradual appearance of modular and non-modular loop prototypes reveals the early rise of NM loops in evolution.

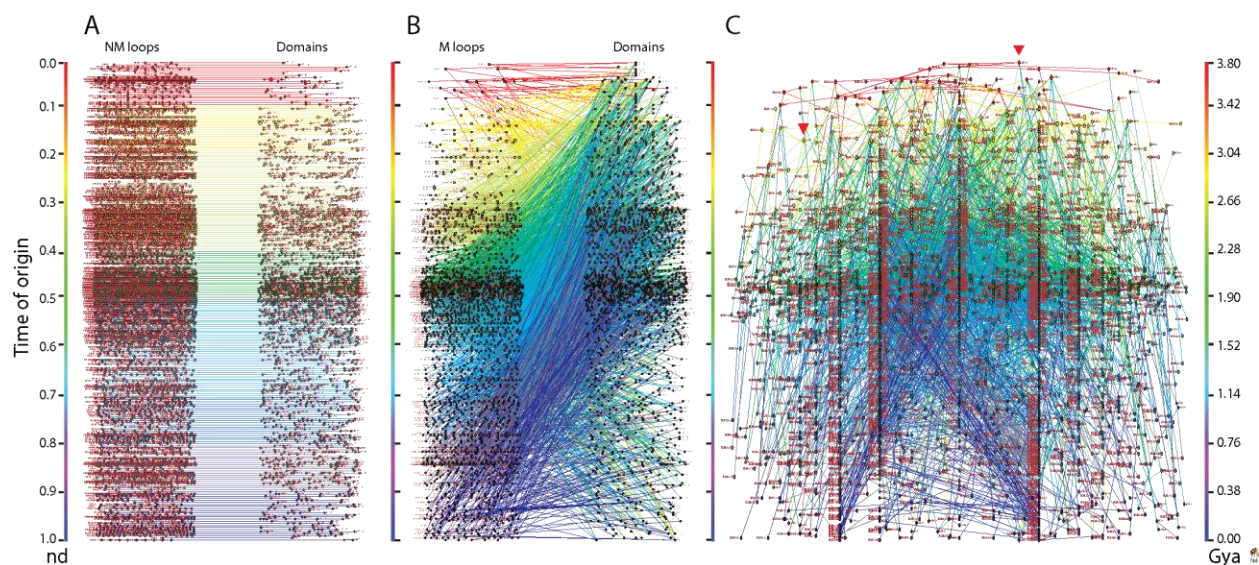


Figure 3.3. Bipartite networks describe the origin and evolution of structural domains by recruitment of non-modular (NM) and modular (M) loop prototypes in protein evolution. Networks uncover how domains share NM prototypes (A) or M prototypes (B and C) along the evolutionary timeline. Loop and domain nodes are colored according to time events, labeled using established ArchDB and SCOP nomenclature, respectively, and arranged top-down according to time of origin (age, *nd*) displayed on a relative 0-to-1 scale or on a ‘billions of years ago’ (Gya) scale time-calibrated with a molecular clock of domains. The network linking M prototypes to domains in bipartite (B) and time event waterfall format (C) represents an evolving ‘elementary functionome’ (EF) describing the recruitment of protein loop modules. Nodes were scaled proportional to their weighted degree, i.e. the sum of the weights of all edges of the nodes. Prototype hits to structural domains in proteomes were not used to weight edges to avoid complication in interpretation of weighted network projections. Red arrowheads indicate the origin of major waves of recruitment in the waterfall network. The horizontal expansion is dictated by VOS clustering, which elucidates formation of modules along the evolutionary timeline (see methods).

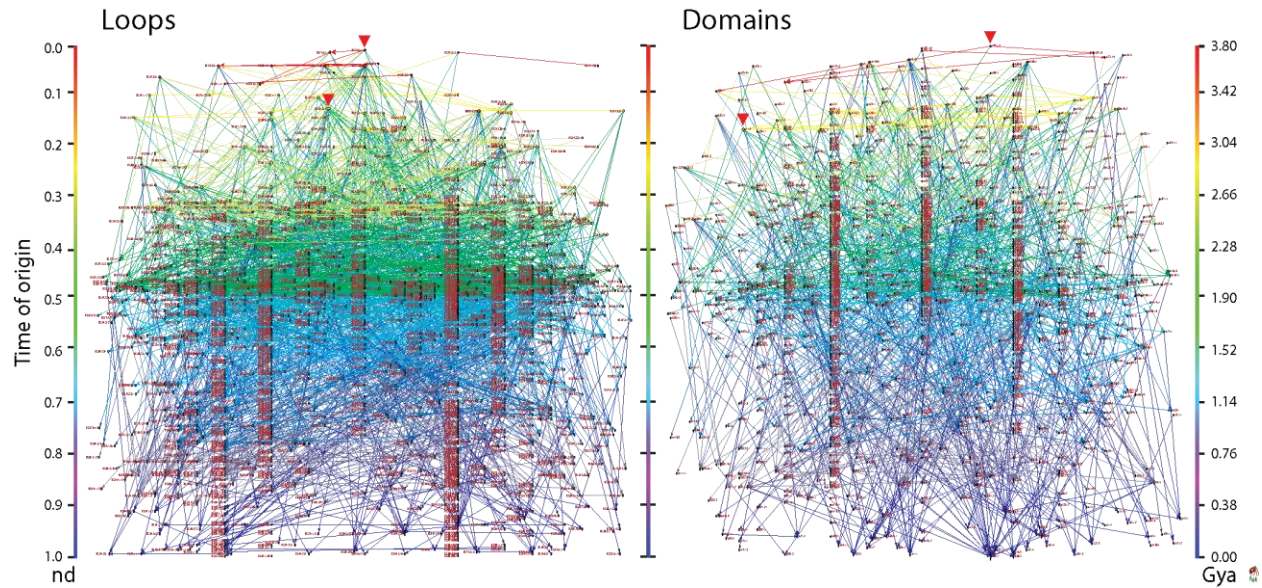


Figure 3.4. EF network projections in waterfall layout. (A) Loop network defined by 1937 prototypes (ellipsoids) and arc connections (arrows) representing sharing of domain structures. **(B)** Domain network defined by 1442 domains (rhomboids) and arc connections representing sharing of loop prototypes. Loop and domain nodes in their uni-modal graph representations were labeled using established ArchDB and SCOP nomenclature, respectively, arranged top-down in the order of time events, and colored according to age (*nd*) on a relative 0-to-1 scale or on a ‘billions of years ago’ (Gya) scale. Links (directed) were colored by the age of a destination node. The 2-dimensional scale of nodes was kept proportional to their weighted degree. In particular, the horizontal and vertical sizes of the loop and domain symbols were made proportional to the weighted outdegree and indegree, respectively, showcasing source-and-sink relationships. All weighted degree vectors were shifted by a value of 10 to avoid vanishing of 0-degree entities. The width of arcs joining the loops and domains was made proportional to the number of shared domains and loops, respectively. Red arrowheads indicate the origin of major waves of recruitment in the time event waterfall. The arcs symbolize the flow of time (random direction for contemporary nodes).

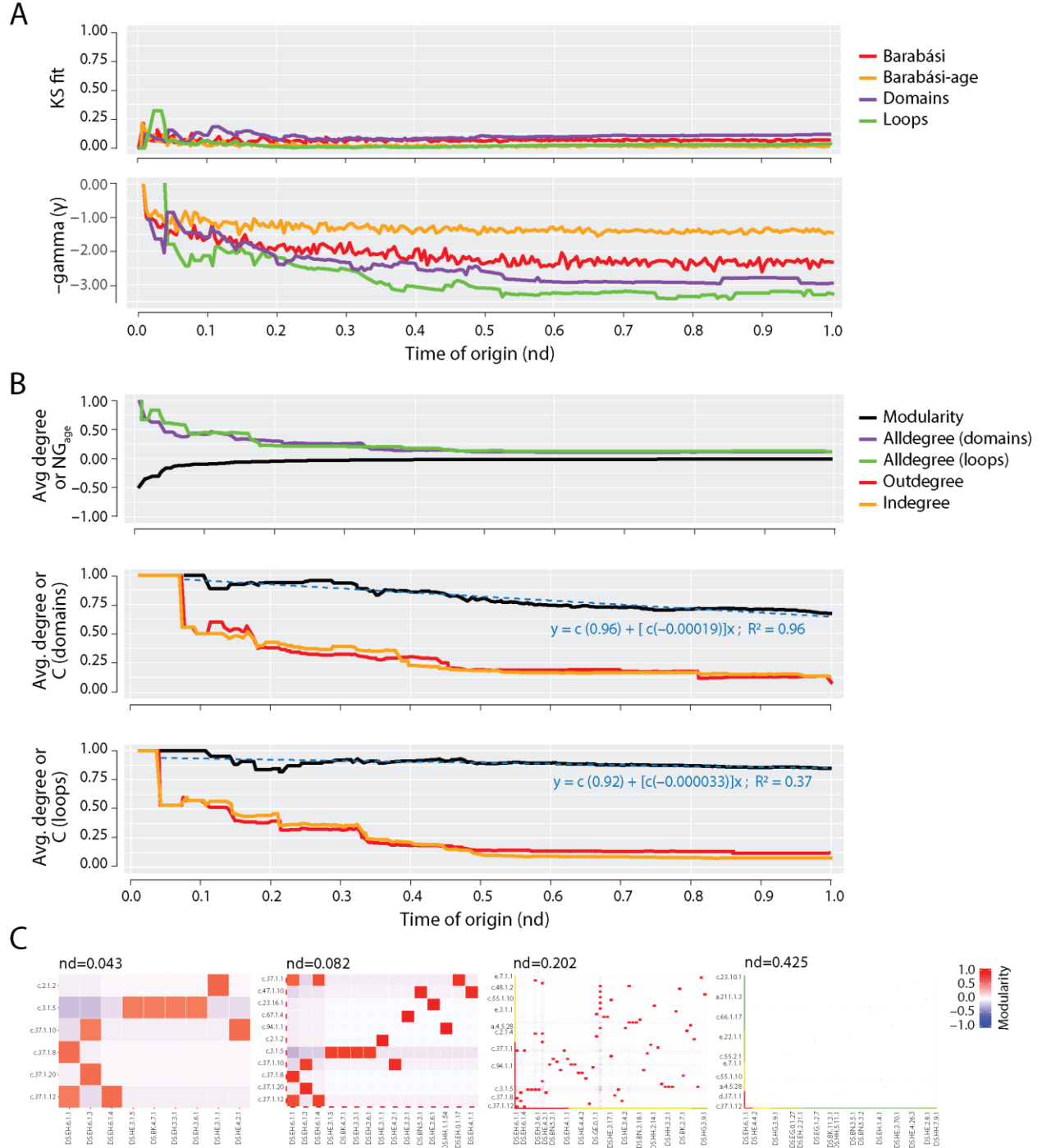


Figure 3.5. Emergence of scale-free and modular behavior in the evolving EF network. (A) Transfer of the scale-free property in the EF network. The KS fit statistic measures network degree deviations from the fitted power law distribution. Lower KS values indicate better fit. The reference Barabási (red) and Barabási-age (orange) curves are included for comparison. The generated scale-free network controls consider the preferential attachment probability of an old node to be proportional to its degree (Barabási) or to both its age and degree (Barabási-age). Power law decay γ exponents of the lower panel ‘scale-free’ levels of heterogeneity in networks, with $\gamma > 2$ describing typical heavy-tailed distribution of connectivity. **(B)** Modularity of growing networks. NG with default membership

Figure 3.5 (cont.)

(partition) defined by age (NG_{age}) was computed for the EF network. NG_{age} indicates mixing of nodes by age in an assortative (≥ 0) or disassortative (< 0) manner across modules (Newman & Girvan, 2004). The average *Clustering Coefficient* (C) for loop and domain networks describes the averaged ratio of the triangles to the connected triples over all nodes, where the networks are simplified (undirected/unweighted) (Wasserman and Faust, 1994; Ravasz *et al.*, 2002; Barrat *et al.*, 2004). We report the coefficients of linear regression models (blue lines) over C for the domain network as -0.00019 by network size (N) and -0.350 by age, and those for the loop network as -0.000033 by N and -0.091 by age. Linear regression lines shown are by N . Normalized average degree (avg. degree) curves, computed as mean-/ max-degree of the network at an event, were included as reference controls. Separate curves were computed for the ‘alldegree’ of loop and domain portions of the EF network and for the ‘outdegree’ and ‘indegree’ of loop and domain networks. Degrees were cumulative and weighted. Scores and indices were calculated for each event of the evolving networks. Time of origin (nd) is indicated in a relative 0-to-1 scale. (C) Progression of pairwise modularity in the EF network. The cells of the heatmaps represent modular strength between a loop and domain as compared to their individual connectivity with the rest of the network, scaled by the absolute $\log_{10}$ value of the network wide modularity index NG_{age} at that event (Newman & Girvan, 2004). The first three panels illustrate the hidden switch of power law and modularity properties between loops and domains. The last panel corresponds to the most distinguishable plateaued EF network. The significant loop prototypes and domain structures involved in the two major waves of functional innovation along with sister hubs having red tiles are displayed using established ArchDB and SCOP nomenclature, ordered ascendingly and color-coded according to node age.

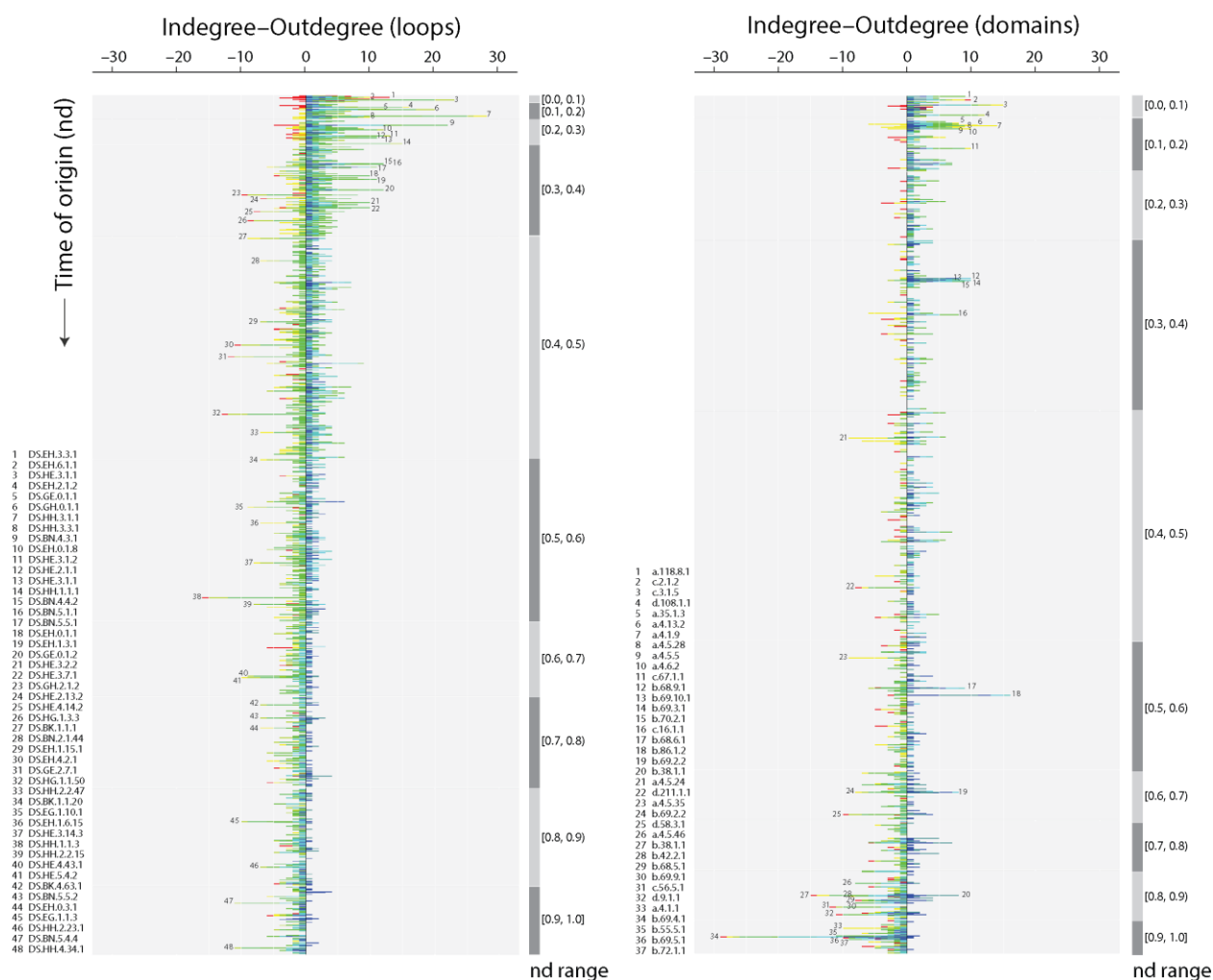


Figure 3.6. Chronological accumulation of connectivity in loop and domain networks. The stacked bar charts depict the chronological accumulation of connections (arcs) in the loop and domain networks along time events of the timeline. Nodes in 99th percentile of connectivity are labeled using ArchDB and SCOP nomenclature. Each event corresponds to the discovery of loops and domains from one of 206 and 226 events, respectively, along a timeline that spans the origin of proteins ($nd = 0$) and the present ($nd = 1$). For visualization purposes, the timeline of events was coarse-grained into 10 age bins. For each node, the number of connections to nodes appearing earlier (indegree) or later (outdegree) in evolution were recorded and displayed as colored stacks in the stacked bars colored red-to-blue following time. The charts portray sink-source relationships in the recruitment of elementary functions viewed from the perspectives of loops and domains.

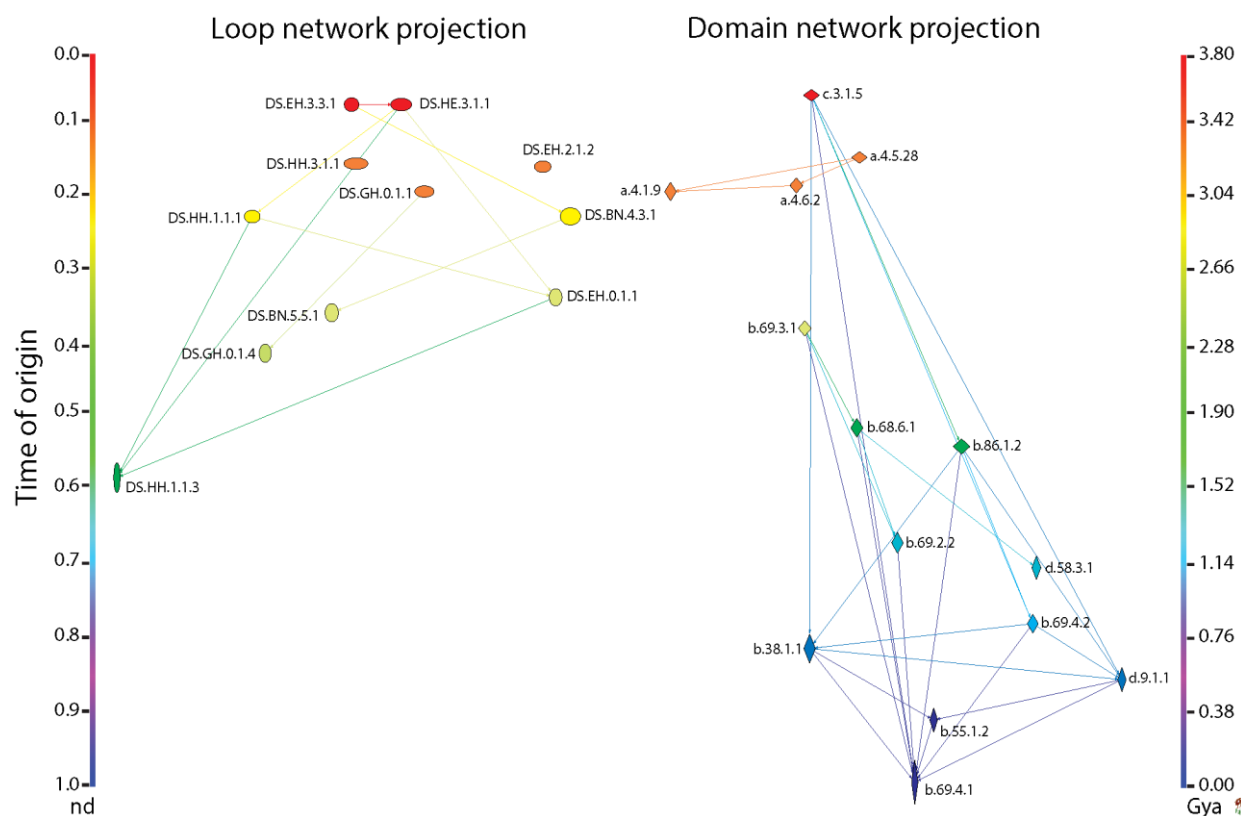


Figure 3.7. EF network projections in waterfall layout describing the evolution of loops and domains with the largest (100th percentile) network connectivity. The loop and domain network projections of 1,442 and 1,937 nodes, respectively, were each reduced with the restrictive criterion of excluding nodes with combined outdegrees and indegrees $\leq 99\%$ of those of the rest of the nodes. The set of arcs (arched arrows symbolizing flow of time) in each network was also reduced to pairing events between nodes with 100th percentile connectivity and excluded those between contemporary nodes. Nodes are arranged top-down and colored according to age (*nd*) on a relative 0-to-1 scale or on a ‘billions of years ago’ (Gya) scale that describes evolutionary time events. Arcs are color-coded according to the age of the impending or more recent of the joined nodes. Loop and domain nodes were labeled with ArchDB and SCOP descriptors, respectively. To showcase source-and-sink relationships, node symbol sizes were scaled proportional to the weighted outdegree and indegree along the horizontal and vertical axes, respectively. Weighted degrees were shifted by a value of 10 to include 0-degree nodes for better visualization. The modular spread of nodes was based on VOS clustering.

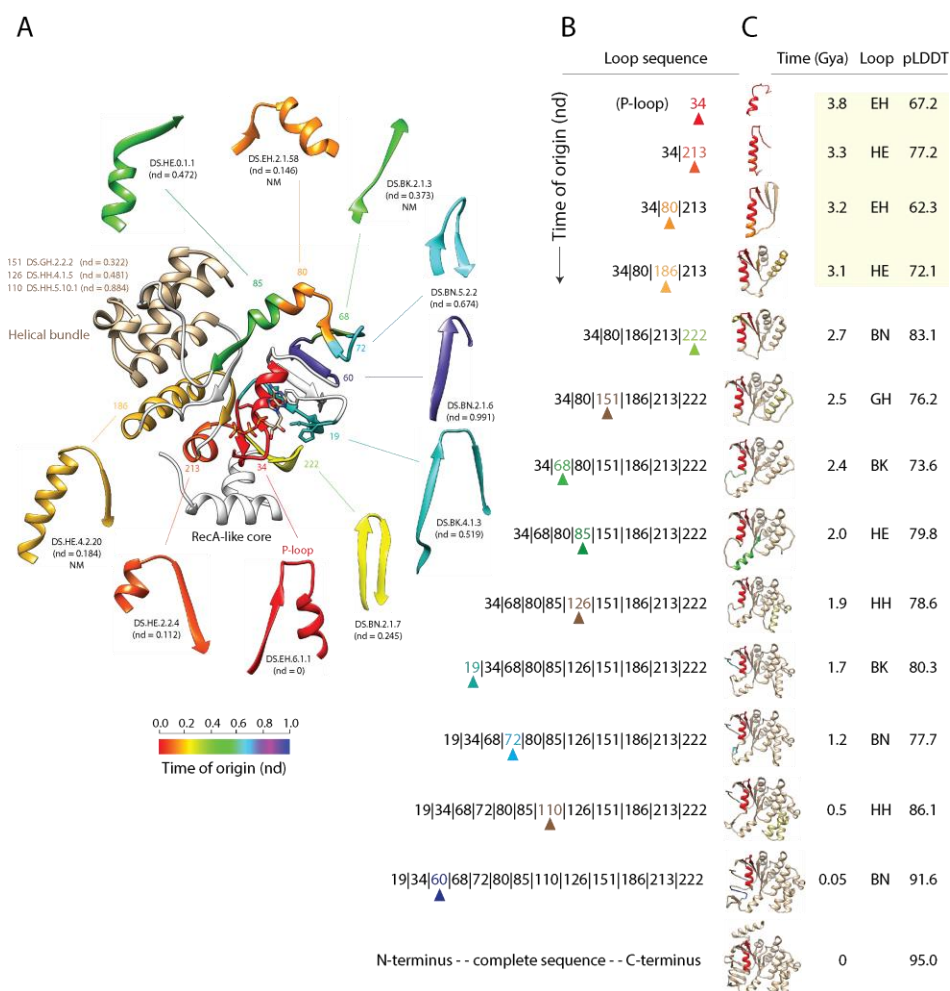


Figure 3.8. Tracing the evolutionary history of loop prototypes embedded in the structure of the most ancient structural domain. (A) A crystallographic model describing the atomic structure of the P-loop containing ATP-binding domain of the histidine permease from *Salmonella typhimurium* (PDB entry 1B0U) shows a nucleotide-binding RecA-like core and a helical bundle region with a signature Mg-ATP binding site that confers substrate specificity. The different loop prototypes that make up the RecA-like core domain structure are colored according to their time of origin, which is given as relative age (nd) in a scale from 0 (origin of proteins) to 1 (the present). They are labeled according to ArchDB classification and mapped to loop structures labeled with numbers describing their position in the sequence relative to the N-terminal amino acid. For simplicity, loop structure numbers and prototypes of the helical bundle are not traced but are listed. (B) A time-ordered series of growing molecules was constructed by stitching loops together according to their time of origin. The sequence of loops is given from N- to C-terminus, with loops labeled in numbers and stitching interfaces indicated by pipe symbols. The last loop to be added to the sequence is indicated with an arrowhead and colored according to its age. (C) 3-dimensional atomic structures of the growing molecules were modeled directly from their sequences with AlphaFold2. The age of the first loop (the P-loop) and the last loop to be added to the structure are colored in the growing structures. The time of origin (Gya), makeup of bracing secondary structures of the loop, and pLDDT confidence level of the *ab initio* prediction are given for each growing molecule.

3.7 References

1. G. Caetano-Anollés, M. Wang, D. Caetano-Anollés, Structural phylogenomics retrodicts the origin of the genetic code and uncovers the evolutionary impact of protein flexibility. *PLoS One* **8(8)**, e72225 (2013).
2. E.N. Trifonov, Z.M. Frenkel, Evolution of protein modularity. *Curr. Op. Struct. Biol.* **18**, 335-340 (2009).
3. Y. Sobolevsky, R.C. Guimaraes, E.N. Trifonov, Towards functional repertoire of the earliest proteins. *J. Biomol. Struct. Dyn.* **31(11)**, 1293-1300 (2013).
4. J. Söding, A.N. Lupas, More than the sum of their parts: on the evolution of proteins from peptides, *Bioessays* **25(9)**, 837–846 (2003).
5. E. Papaleo, G. Saladino, M. Lambrughì *et al.*, The role of protein loops and linkers in conformational dynamics and allostery', *Chemical Rev.* **116(11)**, 6391–6423 (2016).
6. J.F. Leszczynski, G.D. Rose, Loops in globular proteins: a novel category of secondary structure, *Science* **234(4778)**, pp. 849–855 (1986).
7. I.N. Berezovsky, E.N. Trifonov, Van der Waals locks: loop-n-lock structure of globular proteins. *J. Mol. Biol.* **307**, 1419–1426 (2001).
8. E. Aharonovsky, E.N. Trifonov, Protein sequence modules. *J. Biomol. Struct. Dyn.* **23(3)**, 237-242 (2005).
9. I.N. Berezovsky, A.Y. Grosberg, E.N. Trifonov, Closed loops of nearly standard size: common basic element of protein structure. *FEBS Lett.* **466**, 283–286 (2000).
10. A. Goncarenco, I.N. Berezovsky, Prototypes of elementary functional loops unravel evolutionary connections between protein functions. *Bioinformatics* **26**, i497-i503 (2010).
11. A. Goncarenco, I.N. Berezovsky, Exploring the evolution of protein function in Archaea. *BMC Evol. Biol.* **12(1)**, 75 (2012).
12. A. Goncarenco, I.N. Berezovsky, Protein function from its emergence to diversity in contemporary proteins, *Physical Biol.* **12(4)**, 45002 (2015).
13. I.N. Berezovsky, E. Guarnera, Z. Zheng, Basic units of protein structure, folding, and function', *Prog. Biophys. Mol. Biol.* **128**, 85–99 (2017).
14. M.F. Aziz, K. Caetano-Anollés, G. Caetano-Anollés, The early history and emergence of molecular functions and modular scale-free network behavior. *Sci. Rep.* **6**, 25058 (2016).
15. V. Alva, J. Söding, A.N. Lupas, A vocabulary of ancient peptides at the origin of folded proteins. *eLife* **4**, e09410 (2015).
16. S. Nepomnyachiy, N. Ben-Tal, R. Kolodny, Global view of the protein universe. *Proc. Natl. Acad. Sci. USA* **111(32)**, 11691–11696 (2014).

17. S. Nepomnyachiy, N. Ben-Tal, R. Kolodny, Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. *Proc. Natl. Acad. Sci. USA* **114(44)**, 11703–11708 (2017).
18. G. Caetano-Anollés, M.F. Aziz, F. Mughal, D. Caetano-Anollés, Tracing protein and proteome history with chronologies and networks: folding recapitulates evolution. *Exp. Rev. Proteomics* **18(10)**, 863-880 (2021).
19. G. Caetano-Anolles, M. Wang, M., D. Caetano-Anolles, J.E. Mittenthal, The origin, evolution and structure of the protein world. *Biochem. J.* **417**, 621–637 (2009).
20. G. Caetano-Anollés, D. Caetano-Anollés, An evolutionarily structured universe of protein architecture. *Genome Res.* **13(7)**, 1563–1571 (2003).
21. H. Edwards, S. Abeln, C.M. Deane, Exploring fold space preferences of new-born and ancient protein superfamilies. *PLoS Comput. Biol.* **9(11)**, 1003325 (2013).
22. M. Wang, G. Caetano-Anollés, The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. *Structure* **17(1)**, 66-78 (2009).
23. K.M. Kim, G. Caetano-Anollés, Emergence and evolution of modern molecular functions inferred from phylogenomic analysis of ontological data. *Mol. Biol. Evol.* **27(7)**, 1710-1733 (2010).
24. I. Koç, G. Caetano-Anollés, The natural history of molecular functions inferred from an extensive phylogenomic analysis of gene ontology data. *PLoS One* **12(5)**, e0176129 (2017).
25. N. Nath, J.B. Mitchell, G. Caetano-Anollés, The natural history of biocatalytic mechanisms. *PLoS Comput. Biol.* **10(5)**, e1003642 (2014).
26. C. Debès, M. Wang, G. Caetano-Anollés, F. Gräter, Evolutionary optimization of protein folding. *PLoS Comput. Biol.* **9(1)**, e1002861 (2013).
27. J. Bonet, J. Planas-Iglesias, J. Garcia-Garcia, M.A. Marín-López, N. Fernandez-Fuentes, B. Oliva, ArchDB 2014: structural classification of loops in proteins. *Nucleic Acids Res.* **42(D1)**, D315--D319 (2014).
28. A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247(4)**, 536-540 (1995).
29. F.R.K. Chung, P. Erdős, J. Spencer, On the decomposition of graphs into complete bipartite subgraphs. In: P. Erdős, L. Alpar, G. Halasz, A. Saeközy, eds. *Studies in pure mathematics, To the memory of Paul Turán* (Birkhäuser Verlag, Budapest 1983), p95-101.
30. M. Wang, M., Y.-Y. Jiang , K.M. Kim, G. Qu, H.-F. Ji, J.E. Mittenthal, H.-Y. Zhang, 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.* **28**, 567-582 (2011).

31. G. Caetano-Anollés, K.M. Kim, D. Caetano-Anollés, The phylogenomic roots of modern biochemistry: Origins of proteins, cofactors and protein biosynthesis. *J. Mol. Evol.* **74**, 1-34 (2012).
32. K. Caetano-Anollés, G. Caetano-Anollés, Structural phylogenomics reveals gradual evolutionary replacement of abiotic chemistries by protein enzymes in purine metabolism. *PLoS One* **8(3)**, e59300 (2013).
33. R. Diestel, *Graph theory. Graduate Texts in Mathematics*, Vol. 173, Fourth Edition (Springer-Verlag, Heidelberg, 2010).
34. W. Delaney, E. Vaccari, *Dynamic models and discrete event simulation* (Marcel Dekker Inc., New York, 1989).
35. M.H. MacDougall, *Simulating computer systems: Techniques and tools* (MIT Press, Cambridge, 1987).
36. M. Pidd, *Computer simulation in management science* (John Wiley & Sons, New York, 2004).
37. D.S. Tawfik, Messy biology and the origins of evolutionary innovations. *Nature Chem. Biol.* **6**, 692-696 (2010).
38. K.K. Turoverov, I.M. Kuznetsova, A.V. Fonin, A.L. Darling, B.Y. Zalavsky, U.V. Uversky, Stochasticity of biological soft matter: Emerging concepts in intrinsically disordered proteins and biological phase separation. *Trends Biochem. Sci.* **44(8)**, 716-728 (2019).
39. G. Boël, O. Danot, V. de Lorenzo, A. Danchin, Omnipresent Maxwell's demons orchestrate information management in living cells. *Microbial Biotechnol.* **12(2)**, 210-242 (2019).
40. N.J. Van Eck, L. Waltman, VOS: a new method for visualizing similarities between objects. In: H.-J. Lenz H-J, Decker R, editors. *Advances in Data Analysis: Proceedings of the 30th Annual Conference of the German Classification Society* (Springer Verlag, Heidelberg, 2007), 299-306.
41. L. Waltman, N.J. Van Eck, E.C. Noyons, A unified approach to mapping and clustering of bibliometric networks. *J. Informetrics* **4(4)**, 629-635 (2010).
42. T. Kamada, S. Kawai, An algorithm for drawing general undirected graphs. *Inf. Proc. Lett.* **31(1)**, 7-15 (1989).
43. S.H. Strogatz, Exploring complex networks. *Nature* **410**, 268-276 (2001).
44. H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.L. Barabási, The large-scale organization of metabolic networks. *Nature* **407(6804)**, 651-654 (2000).
45. M.E. Newman, Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* **46(5)**, 323-351 (2005).

46. A. Clauset, C.R. Shalizi, M.E. Newman, Power-law distributions in empirical data. *SIAM Rev.* **51(4)**, 661-703 (2009).
47. E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.L. Barabási, Hierarchical organization of modularity in metabolic networks. *Science* **297(5586)**, 1551-1555 (2002).
48. J. Mittenenthal, D. Caetano-Anollés, G. Caetano-Anollés, Biphasic patterns of diversification and the emergence of modules. *Front. Genetics* **3**, 147 (2012).
49. M.E. Newman, M. Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **69(2)**, 026113 (2004).
50. A. Wagner, D.A. Fell, The small world inside large metabolic networks. *Proc. Roy. Soc. London Series B: Biol. Sci.* **268(1478)**, 1803-1810 (2001).
51. A. Barrat, M. Barthélemy, R. Pastor-Satorras, A. Vespignani, The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. USA* **101(11)**, 3747-3752 (2004).
52. S. Wasserman, K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge University Press, Cambridge, 1994).
53. D.J. Watts, S.H. Strogatz, Collective dynamics of ‘small-world’ networks. *Nature* **393**, 440–442 (1998).
54. R. Albert, A.L. Barabási, Statistical mechanics of complex networks. *Rev. Modern Phys.* **74(1)**, 47 (2002).
55. F. Mughal, G. Caetano-Anollés, MANET 3.0: Hierarchy and modularity in evolving metabolic networks. *PLoS One* **14(10)**, e0224201 (2019).
56. M.E. Newman, S.H. Strogatz, D.J. Watts, Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64(2)**, 026118 (2001).
57. A. Clauset, M.E. Newman, C. Moore, Finding community structure in very large networks. *Phys. Rev. E* **70(6)**, 066111 (2004).
58. H.S. Kim, J.E. Mittenenthal, G. Caetano-Anollés, Widespread recruitment of ancient domain structures in modern enzymes during metabolic evolution. *J. Integr. Bioinform.* **10(1)**, 214 (2013).
59. M.F. Aziz, G. Caetano-Anollés G. Evolution of networks of protein domain organization. *Sci. Rep.* **11(1)**:12075 (2021).
60. G. Caetano-Anollés, H.S. Kim, J.E. Mittenenthal, The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proc. Natl. Acad. Sci. USA* **104**, 9358–63 (2007).
61. M. Teichmann, H. Dumay-Odelot, S. Fribourg, Structural and functional aspects of the winged-helix domains at the core of transcription initiation complexes. *Transcription* **3**, 1 (2012).

62. G. Caetano-Anollés, L.S. Yafremava, H. Gee, D. Caetano-Anollés, H.S. Kim, J.E. Mittenhal, The origin and evolution of modern metabolism. *Intl. J. Biochem. Cell. Biol.* **41**, 285–297 (2009).
- Barabási, A.L. & Albert, R. Emergence of scaling in random networks. *Science* **286(5439)**, 509-512 (1999).
- Batagelj, V. & Mrvar, A. Pajek-program for large network analysis. *Connections* **21(2)**: 47-57. Website: Batagelj V, Mrvar A, Pajek-program for large network analysis, (1998), <http://vlado.fmf.uni-lj.si/pub/networks/pajek/>, Date of access: 09/30/2015.
- J. Bonet, A. Fiser, B. Oliva, N. Fernandez-Fuentes, S motifs as structural local descriptors of supersecondary elements: classification, completeness and applications. *Bio-Algorithms Med-Systems* **10(4)**, 195-212 (2014).
- J. Bonet, J. Segura, J. Planas-Iglesias, B. Oliva, N. Fernandez-Fuentes, Frag'r'Us: knowledge-based sampling of protein backbone conformations for de novo structure-based protein design. *Bioinformatics* **30**, 1935-1936 (2014).
- Borg, I. & Groenen, P. Modern multidimensional scaling: theory and applications. *J. Educ. Measurement* **40(3)**, 277-280 (2003).
- G. Caetano-Anollés, M.J. Seufferheld, The coevolutionary roots of biochemistry and cellular organization challenge the RNA world paradigm. *J. Mol. Microbiol. Biotechnol.* **23**, 152-177 (2013).
- Csardi, G. & Nepusz, T. The igraph software package for complex network research. *Intl. J. Complex Systems* **1695(5)**, 1-9 (2006).
- Eddy, S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195. <https://doi.org/10.1371/journal.pcbi.1002195>
- N. Fernandez-Fuentes, J.M. Dybas, A. Fiser, Structural characteristics of novel protein folds. *PLoS Comput. Biol.* **6**, e1000750 (2010).
- Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313**, 903–19. <https://doi.org/10.1006/jmbi.2001.5080>
- A. Harish, G. Caetano-Anollés, Ribosomal history reveals origins of modern protein synthesis. *PLoS One* **7(3)**, e32776 (2012).
- Ihaka, R. & Gentleman, R. R: A language for data analysis and graphics. *J. Comput. Graph Stat.* **5(3)**, 299-314 (1996).
- J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinneger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O.

- Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
- M. Mirdita, K. Schülte, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, ColabFold: making protein folding accessible to all. *Nature methods* **19**, 679-682 (2022).
- Kolaczowski, B., Thornton, J. W. (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431**, 980–984.
- A. Kryzhtafovych, T. Schwede, M. Topf, K. Fidelis, J. Moult, Critical assessment of methods of protein structure prediction (CASP): round XIV. *Proteins* **89**, 1607–1617 (2021).
- V. Mariani, M. Biasini, A. Barbato, T. Schwede, IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**(21), 2722-2728 (2013).
- Murtagh, F. & Legendre, P. Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *J. Classification* **31**(3), 274-295 (2014).
- Nasir, A., and Caetano-Anollés, G. (2015) A phylogenomic data-driven exploration of viral origins and evolution. *Science Advances* **1**, e1500527.
- O'Leary, N.A., Wright, M.W., Brister, J.R., et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Reseach* **44**, D733-D745.
- E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin, UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**(13), 1605-12 (2004).
- Pons, P. & Latapy, M. Computing communities in large networks using random walks. *J. Graph Algorithms Appl.* **10**(2), 191-218 (2006).
- R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Website: R Core Team, The R Project for Statistical Computing, (2014), <http://www.R-project.org/>, Date of access: 09/30/2015.
- M.L. Romero Romero, F. Yang, Y-R, Lin, A. Toth-Petroczy, I.N. Berezovsky, A. Goncarenko, W. Yang, A. Wellner, F. Kumar-Deshmukh, M. Sharon, D. Baker, G. Varani, D.S. Tawfik, Simple yet functional phosphate-loop proteins. *Proc. Natl Acad. Sci. USA* **115**, E11943-E11950.
- Swofford, D. L. (2022) *Phylogenomic Analysis Using Parsimony and Other Programs (PAUP*)* Ver 4.0b10 (Sinauer, Sunderland, MA).
- Zhang, H.-Y., Qin, T., Jiang, Y.-Y. & Caetano-Anollés, G. Structural phylogenomics uncovers the early and concurrent origins of cysteine biosynthesis and iron-sulfur proteins. *J. Biomol. Struct. Dyn.* **30**(5), 542-545 (2012).
- R. Overbeek, et al. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* **28**(1), 123-125 (2000).

J. Skolnick, H. Zhou, M. Brylinski, Further evidence for the likely completeness of the library of solved single domain protein structures. *J. Phys. Chem. B* **116**, 6654–6664 (2012).

Skolnick J; Gao M. The role of local versus nonlocal physicochemical restraints in determining protein native structure. *Curr. Opin. Struct. Biol* 2021, 68, 1–8 (2021).

Skolnick J; Gao M., Zhou H, Singh, S. AlphaFold2: Why it works and its implications for understanding the relationships of prtein sequence, structure, and function. *J.Chem.Inf.Model.* 61(10), 4827-4831 (2021).

CHAPTER 4: A NETWORK VIEW OF DIPEPTIDE EVOLUTION

4.1 Abstract

In protein evolution, structural, functional, and evolutionary units of proteins must emerge from prior structural states by fulfilling the evolutionary principle of spatiotemporal continuity. Here, we show that these prior states involve a combinatorial origami of dipeptides capable of forming flexible protein domain structures. The interactome of 400 dipeptides with 1,475 SCOP fold families (domains) and 1,561 proteomes was examined for modular and scale-free properties. This bipartite network and the resulting dipeptide projections helped examine how dipeptides behave as building blocks of domains while forming a historical ‘dipeptide footprint’ (D-F). Ages of dipeptides were directly borrowed from the associated domains. We tested the hypotheses of whether collective dipeptide frequencies can be significantly modeled as the proteome size. We also tested if the collective normalized dipeptide frequencies can be significantly modeled as dipeptide age. We also examined how dipeptide abundances relate to proteome size and dipeptide age in factorial design. Finally, we investigate if the domain and proteome-sharing based dipeptide networks follow a power-law. Most properties were significant and rejected the null hypotheses with networks turning out to be scale-rich instead of scale-free, following a hierarchically modular paradigm. A study on the randomness of the dipeptide networks and further exploration of recruitment waves is recommended to dissect hidden network properties.

Index Terms: Dipeptide, Fold Families, evolutionary age, network, connectivity, scale-free, scale-rich, modularity.

4.2 Introduction

"... Does this Aleph exist in the heart of a stone? Did I see it there in the cellar when I saw all things, and have I now forgotten it? Our minds are porous and forgetfulness seeps in; I myself am distorting and losing, under the wearing away of the years, the face of Beatriz..."— Jorge Luis Borges, The Aleph and Other Stories

Amino acid and dipeptide frequencies in proteomes have been used to retrodict the origins of the genetic code (Caetano-Anollés, Wang and Caetano-Anollés, 2013). Prebiotic chemistry is a fascinating field aiming to discover the cosmic origin of life, which seems to have begun by dealing with simple molecules, whose composition quickly gained complexity with oligomerization (Mayer, 2022). From a set of 20 monomers representing the 20 proteinogenic amino acids, 400 dimers are expected, extendable to 3,200,000 pentamers, or $> 10^{13}$ decamers. These “messy” starting conditions hold enormous power of potentially functional oligomers. However, a “selecting” structure such as a multilayered membrane or a vesicle microenvironment, may choose and grow those molecules from the existing pool to fulfill essential functions (Aziz, Caetano-Anollés and Caetano-Anollés, 2016; Aziz and Caetano-Anollés, 2021; Caetano-Anollés *et al.*, 2021). But, if such a “selector” has to also perpetually go through a selection process (perhaps in a hierarchically modular manner (Ravasz *et al.*, 2002)), the accumulated oligomers may be simplified and combined to fulfill more complex functions (Wang, Kurland and Caetano-Anollés, 2011), which eventually improves the adaptability of the “selectors”. Thus, after multiple steps of evolutionary optimization, a functional unit of reasonable order and complexity would be formed. This apparently random process of multi selection steps has been a subject of many investigations (Mayer, 2022). In this study, we attempt to enumerate the evolutionary roles and contexts of these basic dimers, which we will call dipeptide (dipeptide) units, trying to summarize a framework of

their contribution to protein domain structural makeup over billions of years (Gy) of evolution (Wang *et al.*, 2011). Evolutionary heat maps of dipeptide abundances in structural domains defined at fold family (domain) level of the Structural Classification of Proteins (SCOP) (Murzin *et al.*, 1995) ordered by domain age give a glimpse of dipeptide compositions of domain structures and their evolution. A unique aspect of this dipeptide frequency distribution among domains is the age association of domains, which implies an evolutionary timeline of architectural development (Caetano-Anollés, Wang and Caetano-Anollés, 2013). The non-random hourglass like patterns of dipeptide compositions unfolding in time uncover unique signatures of dipeptide usage in domains. Our goal was to mine these signatures to understand any evolutionary link that may exist between dipeptides and domain makeup. In this study, we first bring data sets in a form of networks using a graph theoretical approach. For that purpose, we constructed bipartite networks with nodes representing either dipeptides or domains or dipeptides and proteomes of the organismal species analyzed, respectively. Links in these networks represent dipeptide or proteome sharing at certain level of dipeptide abundance. We then analyze network properties (e.g., degree distribution) as well as statistical patterns in the initial data sets using statistical techniques such as multivariate analysis of variance (MANOVA), principal components analysis (PCA) and generalized linear models (GLMs). For the initial data set, for instance, the matrix of dipeptide abundances was modeled in MANOVA against the sizes of proteomes, and against the age of domains. Principal components covering the most variance among the 400 dipeptides abundances were also identified. In GLMs, the data sets were melted to align abundances as categories of dipeptides and then the vectors of abundances were modeled as response variables to the main effects of proteome size and age, respectively.

4.3 Results and Discussion

Bipartite networks were generated from abundance data sets of domains and species-based dipeptides to analyze connectivity patterns between dipeptides and domains, and between dipeptides and proteomes. A visual and statistical analysis of these networks was conducted based on dipeptides network's degree (i.e., number of links of each node). Multivariate analysis using MANOVA (Multivariate Analysis of Variance), PCA (Principal Components Analysis), and GLMs (Generalized Linear Models) was also included to analyze the effects related to dipeptide degree data using Poisson family. Also, the domain and genome-sharing based dipeptide networks were tested for the preferential attachment principle (Barabási and Albert, 1999; Jeong *et al.*, 2000; Clauset, Shalizi and Newman, 2009) and modularity (Clauset, Newman and Moore, 2004). The various analyses of data and networks initially addressed the following questions:

- a) Can the collective dipeptide frequencies be significantly modeled as proteome size?
 - II. How many PCs in the normalized dipeptides abundances vs. species data should be retained in order to explain a high percentage of the variation in the data?
- b) Can the collective normalized dipeptide frequencies be significantly modeled as dipeptide age?
 - II. How many PCs in the normalized dipeptides abundances vs. domains data should be retained in order to explain a high percentage of the variation in the data?
- c) Do dipeptide abundances relate to proteome size in factorial design?
- d) Do dipeptide abundances relate to dipeptides age in factorial design?

- e) Is the domain and genome-sharing based dipeptide network degree explained by the dipeptides?
- f) Are the domain and genome-sharing based dipeptide network significantly scale-free or modular?

4.3.1 The origin and evolution of a ‘dipeptide footprint’ in protein structure elicited from abundance dataset of 400 possible dipeptides and significantly described by the main effects of proteome size and age

We here provide multivariate statistical analyses based on abundances of 400 dipeptides in select proteomes and trace the evolutionary age of domain structures over a bimodal graph-theoretic representation of domains and their associated dipeptide structural units. We decomposed the bipartite representation into one-mode network projections using mathematical properties of finite graphs (Chung, Erdős and Spencer, 1983). The strategy is illustrated in Fig. 4.1. The model bipartite graph can be dissected into two separate sets of vertices (nodes), representing dipeptides (circles) and domains (rhomboids) in this study (Fig. 4.1C, D). Bipartite graphs eventually project shared connectivity within each set of nodes (Diestel, 2010). Inter-links (undirected edges) between nodes of the two sets become the basis of intra-links, i.e., connectivity between nodes in separate networks of each node set. Higher sharing across nodes in the bimodal graph results in stronger connectivity of the extracted separate set (one-mode) graphs. Thus, links in the domain network projection would represent the ‘footprints’ of shared dipeptides in their structural makeup. Similarly, the links in the dipeptide network projection would indicate events when dipeptides joined on the same ‘film’ sequence forming a domain.

Together, this evolutionary bipartite graph and its projections describe the co-option events of ancient dipeptide units carrying structural and functional messages to generate and record subsequent repertoires of domain structures and their associated molecular functions. The bipartite network thus posits the gradual makeup and deep evolution of a ‘dipeptide footprint’ (D-F). The dipeptide units represent one of 400 possible pairs of a vocabulary of 20 amino acids establishing the backbone of the genetic code (Caetano-Anollés, Wang and Caetano-Anollés, 2013). These amino acid pairs provide the most abstract strip of message units translated to cater a host of structural and functional roles. The domains represent bona fide structural, functional and evolutionarily conserved units of proteins (Murzin *et al.*, 1995). When a link is established in the D-F bipartite network between any two members of the individual sets, it indicates the presence of a ‘footprint’ in the structural sequence of an associated domain module. Altogether, these networks describe combinatorial (syntactic) interpretations of a ‘film’ record of dipeptides in which context-dependent rules (pragmatics) determine contextual rules (semantics), much like as in a language processing Turing machine (Caetano-Anollés, Wang and Caetano-Anollés, 2013). Hence, network connectivity describes an evolutionary association of the abstract alphabet of dipeptide units with the structural alphabet of domains.

To elicit evolutionary information from the D-F bimodal network and its domain and dipeptide one-mode projections ages were assigned to its nodes and links (Figs. 4.2 and 4.3). These age mappings make the evolutionary growth of networks in time visible. The ages were retrieved from a structural phylogenomic analysis of 2,397 fold-family (FF) protein domain structures (out of 3,464 defined by the Structural Classification of Proteins (SCOP) 1.73 (Murzin *et al.*, 1995)) was used to derive a timeline from a phylogenomic tree reconstructed from a structural census in the genomes of 420 free-living organisms from all three cellular superkingdoms (FL420) (Caetano-

Anollés, Wang and Caetano-Anollés, 2013). The relative ages based on node-distances (*nd*) of the domain structure were extracted from this timeline. Calibration points of protein domain structures associated with microfossil, fossil and biogeochemical evidence, biomarkers, and first-appearance of clade-specific domains (integrated with molecular, physiological, paleontological, and geochemical data) yielded a timeline of geological ages on a relative time scale (Wang *et al.*, 2011). These ages of domains are proportional to geological time for SCOP fold families (Caetano-Anollés, Kim and Caetano-Anollés, 2012; Caetano-Anollés and Caetano-Anollés, 2013), provided that the fold families are the most ancient in a fold superfamily. Since an ancestral dipeptide unit needs to be present to be introduced into a contemporary or relatively modern protein domain structure, the ages of first appearance of dipeptides were borrowed from ages of either the first appearance of a linked domain or of the second oldest domain in case of dipeptides mapping with different domains present in multiple proteins.

The MANOVA models of “Number of Proteins” and “Number of dimers” and their interaction with the 400-column response matrix of dipeptide abundances in the proteomes of different species rejected the null hypothesis (a) dipeptide frequencies cannot be modeled as proteome size; and showed significance of the two effects and their interaction (Supplementary Fig. S1A), with $p\text{-value} < 2.2\text{e-}16$. “ND_val” or age was modeled as main effect of the matrix of dipeptides abundances in domains, and the response was significantly described by the age of domains with $p\text{-value } 3.561\text{e-}07$, rejecting the null hypothesis b) normalized dipeptide frequencies cannot be modeled as dipeptide age (Supplementary Fig. S1B). These results and associated graph plots are well in agreement with the importance of dipeptides in establishing structural innovations of fold families over the evolutionary timeline (Caetano-Anollés, Wang and Caetano-Anollés, 2013).

4.3.2 The variance in the distributions of dipeptide abundances is high, much greater in the domains than in proteomes

The PCA results showed that we needed to retain 152 PCs, with standard deviation greater than one, to explain only up to 68.73% of the variation in dipeptides abundances in the domains (Supplementary Fig. S2A, B), rejecting the null hypothesis b_II) PCs in the normalized dipeptides abundances vs. domains data should not be retained to explain a high percentage of the variation in the data. In the case of proteome data, just 5 PCs, selected by the same criterion of $\text{std.dev} > 1$, were able to explain over 98% of the variance in dipeptides abundances across proteomes (Supplementary Fig. S2B), still rejecting the null hypothesis a_II) PCs in the normalized dipeptides abundances vs. species data should not be retained in order to explain a high percentage of the variation in the data. The scree plots reaffirm higher variance dipeptides abundances introduced in domains. While there is only one inflection point for the PCA plot of proteomes at PC2 (Supplementary Fig. S2C), covering the most variance, the PCA plot of domains shows at least three inflection points, with the major point appearing at PC4, covering just about 70% of the variance (Supplementary Fig. S2D). The plot between PC1 and PC2 of proteome data, capturing highest variance, prominently shows ‘Homo sapiens’ distinctively separate from the cluster (Supplementary Fig. S3A, B). Interestingly, although PC4 of domains data set introduces more variation than PC3 (Supplementary Fig. S2D), indicated by a convex elbow at PC4 after the dip at PC3, the vastest spread was achieved by PC2 and PC3 (Supplementary Fig. S3C, D). It is noteworthy that the middle-aged fold families show highest variance in terms of dipeptides abundance, spreading all around the central cluster.

4.3.3 Early history of dipeptide-based protein makeup is manifested by time event ‘waterfall’ networks and the factorial design of dipeptide abundances as modeled by proteome size and age

A timed network system can be modeled as a discrete sequence of events (network growth at temporal intervals) using Discrete Event Simulation (DES) (MacDougall, 1987; Delaney and Vaccari, 1989; Pidd, 2006). DES typically is a computer-based modeling tool used to analyze complex network behavior and how it affects the interacting entities on consecutive events. A DES supervises a system to stay consistent between events. Eventually, an *incremental step function*, allows flow of time from event to event in a discretized manner. Here, we borrow the DES rationale and map discrete evolutionary ‘time steps’ (events) to the growing structure of an undirected and unweighted D-F bimodal network and its one-mode weighted projections (Figs. 4.2 and 4.3). Events marked by the first appearance of various domains and dipeptides, along with accompanying linkages, generate novel molecular structures and functions. Amino acid pair frequencies yielding abundances of 400 dipeptides in different domains were calculated by assigning structures to sequences from the Protein Data Bank (PDB) (June 20, 2012) (Berman *et al.*, 2000) with the Hidden Markov Models (HMMs) from SUPERFAMILY (Gough *et al.*, 2001). The set of PDB sequences was culled as $\leq 25\%$ using the protein sequence-culling server PISCES (Wang and Dunbrack Jr, 2003) and any sequences with two or more domains or unassigned ranges > 30 amino acids were eliminated from the study. The final PDB sequence set covered 2,384 sequences mapped by 1,475 domains. The evolutionary ages were transferred from the domains to the associated dipeptide structural units. Afterwards, we pruned the mappings to a selected subset representing the highest possible abundance and wide distribution inclusive of all 400 dipeptide structures. The maximum abundance cutoff to include the oldest domain c.37.1.12 in the mappings

was 0.014774 but to accommodate ‘first representation’ of the dipeptide MW with the lowest maximum abundance, the threshold was lowered to 0.010752688. This reduced the number of mappings to 32,624 from 230,833 that were with abundance>0, for the purpose of network analysis. The multivariate and proteome-based analysis was still conducted on 230,833 mappings achieved with the abundance>0 criterion. The emergent D-F bipartite network of 400 dipeptides and 1,475 domains is an undirected graph of 32,624 edges with a *network density* (actual/possible number of edges) of 5.53% [$32624/(1475 \times 400) \times 100$] and a *node average degree* (edges per node) of 34.7989 (± 1.0579 [$45.8086/\sqrt{1875}$]), i.e. D-F components had ~35 interlinks on average. The Visualization of Similarity (VOS) clustering method revealed 73 communities (also known as modules) with a *modularity index* of 0.27 (Van Eck and Waltman, 2007; Waltman, van Eck and Noyons, 2010). The adaptive events of the D-F network were shown through arrangement of color-coded component nodes and edges by age, in a top-down bimodal layout that followed the evolutionary timeline of protein domain structures (Fig. 4.2A). The node sizes are proportional to node connectivity, measured by weighted degree, highlighting the hub-like behavior in the network structure. Also, to better visualize evolutionary network patterns, the VOS clusters (comprising of hubs and their neighbors) were spread horizontally using the energy-optimized Kamada-Kawai (Kamada and Kawai, 1989) ‘free’ and ‘optimize inside clusters only’ methods (Fig. 4.2B). The resulting ‘waterfall’ layout illustrates the structural and functional recruitment across and down the events of the waterfall.

The D-F network projections were visualized in waterfall layouts (Fig. 4.3). The connectivity of these one-mode networks based on intra-connected nodes was made evolutionary explicit with the directionality aspects of *arcs*, arrows symbolizing the flow of time from older to younger nodes. The resulting dipeptide and domain directed graphs had 71,564 and 792,586 arcs each, with

network densities of 44.84% $[71564/(400 \times (400-1)) \times 100]$ and 36.45% $[792586/(1475 \times (1475-1)) \times 100]$, and total node average degrees of 357.82 (± 2.39 $[47.8372/\sqrt{400}]$) and 1074.69 (± 7.25 $[278.53/\sqrt{1475}]$), i.e. domains and dipeptides were shared with other ~ 358 dipeptides and ~ 1075 domains on average, respectively. The un-weighted degrees of the domain and proteome sharing based uncultured dipeptides networks (generated with abundance > 0 criterion) were found to be uniformly distributed at the constant value of 399, meaning that every one of the 400 dipeptides shared at least one domain and proteome with every other dipeptide in that dataset. Both the dipeptide and domain one-mode networks showed visual community structure with 4 and 23 clusters each, and modularity indices of 0.1078 and 0.1177, respectively. The number of outward (outdegree) and inward (indegree) directed connections or arcs specified domain and dipeptide nodes as ‘donors’ (sources) or ‘acceptors’ (sinks) of shared functional dipeptides and domains, respectively. The horizontal and vertical scale of node symbols were made proportional to the weighted outdegree and indegree, respectively. This made the visualization of hubs explicit, e.g., a transition from wide to tall symbols along the flow of time indicated source-sink origination transformations of molecular structure and functions. Overall, these transitions expressed an expected increase in the probability of coopting older dipeptides and domains with time, but also a surprising continual recruitment events among recent molecular entities.

We reverified the two major waves of functional innovation originally identified with metabolites (Caetano-Anollés, Kim and Mitterthal, 2007) and elementary functionomes (Aziz, Caetano-Anollés and Caetano-Anollés, 2016), which were defined by Goncarenco and Berezovsky (Goncarenco and Berezovsky, 2010), in the waterfall diagrams of the D-F network and its projections (Figs. 4.2 and 4.3). These waves that were also recently recognized in protein domains

structural evolution (Aziz and Caetano-Anollés, 2021), exhibit separate origins involving sandwich, barrel and bundle domain structures. We further dissected the early origin of these recruitment pathways arising from ancient ‘*p*-loop (superfamily c.37.1)’ and ‘winged helix (superfamily a.4.5)’ domains by utilizing the versatility of the waterfall visualization in form of highly connected (reduced) subnetworks of the hubs in the dipeptide and domain projections (Fig. 4.4). The realization of these evolutionary patterns with various data sources in parallel is remarkable and strongly supports the historical statements we plan to explore further in future. The first larger wave originated ($nd=0.000$) in the *p*-loop containing nucleoside triphosphate (NTP) hydrolase FF domain (c.37.1.12) (Figs. 4.3A and 4.4A) and its contemporary aminoacidic dipeptides GK, LK and VL (Figs. 4.3B and 4.4B). The c.37.1 fold-superfamily domain group is the most ancient Rossmannoid $\alpha/\beta/\alpha$ -layered domain structure that is most popular in the timeline of domain history (Gustavo Caetano-Anollés, Wang, *et al.*, 2009; Caetano-Anollés and Caetano-Anollés, 2013). The dipeptide network shows that this first wave instigated many massive pathways of dipeptide recruitment, over ~ 2.5 billions of years (Gy) of history. The second wave was traced ~ 3 Gya ($nd=0.343$) in the ‘winged helix’ DNA binding FF domain (a.4.5.42) (Figs. 4.3A and 4.4A), and its 28 connected hub dipeptides VL, EL, dipeptide, ..., DH, RH, VY (Figs. 4.3B and 4.4B). The two waves merge through an ancient most dipeptide VL. Remarkably, these two waves rediscovered by the D-F network and its projections denote the same primordial sandwich $\alpha/\beta/\alpha$ -layered structures, β -barrels and helical bundle structures referred earlier as part of the first 54 historical domains in evolution (Caetano-Anollés, Kim and Caetano-Anollés, 2012). Also, the ‘*p*-loop’ and ‘winged helix’ waves notably embed the first two major gateways of enzymatic recruitment identified earlier in metabolism, the first gateway mediated by the c.37 fold that originated in the energy interconversion pathways of the purine metabolism subnetwork, and

the second mediated by the a.4 fold originating in the subnetwork of porphyrin and chlorophyll metabolism, and the biosynthesis of cofactors (Caetano-Anollés, Kim and Mittenthal, 2007; Gustavo Caetano-Anollés, Yafremava, *et al.*, 2009; Caetano-Anollés and Caetano-Anollés, 2013). Obtaining these congruent results in different structural and evolutionary studies supports the historical statements we here propose.

Using the proteome melted or factorial data, it was uncovered that the response variable of abundances can be modeled by proteome size (Supplementary Fig. S4A), with p-value 2.2×10^{-16} output from the ‘Chi Square’ test of ANOVA. The result rejects the null hypothesis c) dipeptide abundances do not relate to proteome size in factorial design. In a similar fashion, ages of domains significantly modeled the abundances with p-value 3.361×10^{-9} (Supplementary Fig. S4B), rejecting the null hypothesis d) dipeptide abundances do not relate to dipeptides age in factorial design. AIC scores for the main models support these conclusions (Supplementary Fig. S4). The data plots further establish the non-random role of dipeptides in the evolution of protein domain architectures (Caetano-Anollés, Wang and Caetano-Anollés, 2013).

4.3.4 The evolutionary emergence and transformations of molecular functions with dipeptides sharing domains and proteomes at a significantly high degree

Aggregated network visualizations of the structural and functional innovation in the highly connected domain and dipeptide subnetworks showed that the largest hubs appeared early in evolution, however, the recruitment process generating new hubs continued throughout the modern structural world (Fig. 4.4; Supplementary Videos 1 and 2). These patterns coincide with those observed in the evolution of protein domain organization (Aziz and Caetano-Anollés, 2021) and elementary functions (Aziz, Caetano-Anollés and Caetano-Anollés, 2016), where most innovation

happened during the first ~1.8 Gy of protein history (Fig. 4.4). Our investigation using waterfall network layouts arranged 225 unique events along a timeline that spans from the origin of proteins ($nd = 0$) to the modern times ($nd = 1$) (Fig. 4.2). The distribution and connectivity of dipeptide and domain nodes across these events provides information about how progressive, widespread, and recurrent is the combinatorial recruitment process that defines dipeptide-based makeup of domains. Here, we included dipeptide connectivity that maximizes abundance in domains and genomes, and thus eliciting the oldest linkages. The largest hubs holding most of the network connectivity of the of dipeptides and domains apparently appeared very early in protein evolution (Figs. 4.2 and 4.3). Notably, D-F innovation was mostly restricted to peers during the first ~0.8 Gy, and the most rigorous during ~2.5 to ~1.25 Gya period of protein history (Fig. 4.2). While the combinatorial process continued till modern times, the ‘sand dunes of time’ show that the contemporary co-option of ancient dipeptides and domains is prominent (Fig. 4.3). Also, the most acceptors of recruitment events appeared during $nd=[0.1-0.6]$ in dipeptides and $nd=[0.3-0.9]$ in domains. These patterns verify those observed earlier in protein function evolution (Aziz, Caetano-Anollés and Caetano-Anollés, 2016; Aziz and Caetano-Anollés, 2021) and were made quantitative by dissecting the source-sink relationships and evolutionary span of network connectivity with bar plots describing the chronological cumulative connectivity along the timeline (Fig. 4.5). The networks indicate an overwhelming majority of modern donation events, some very recent, with relatively younger ($nd=(0.4-0.9]$) sink nodes being acceptors of ancient donors or source nodes of $nd < 0.2$. Also, sink dipeptides appear to be versatile, progressively drawing innovation from donors spanning the entire timeline. Box-and-whisker plots of cumulative weighted indegree and outdegree over the entire timeline (Supplementary Fig. S5) and scatter plots with linear distribution models of degree totals at $nd=1$ (Supplementary Fig. S6) provided further insight into the patterns

of contraction and expansion of mutually adaptive dipeptide and domain innovation. Notably, domains utilized the repertoire of very ancient donors for their functional and structural growth, with cooption events among modern domains continuing till late in evolution ($nd \geq 0.8$), reconfirming evolutionary patterns of recruitment obtained in the enzymatic analysis of metabolic networks (Kim, Mittenthal and Caetano-Anollés, 2013).

Historically, the progression of molecular novelties involved ligation of small polypeptides or dipeptides with limited ordered structure, followed by the formation of larger peptides harboring stable dipeptides (Caetano-Anollés, Kim and Caetano-Anollés, 2012; Caetano-Anollés, Wang and Caetano-Anollés, 2013), which finally combine to form defined folded topologies typical of small protein domains (Berezovsky and Trifonov, 2001; Aharonovsky and Trifonov, 2005; Goncarenko and Berezovsky, 2010). A continuum of these prior forms is expected according to the principle of spatiotemporal continuity. However, the connectivity patterns we observed expose a more fluid emergence of molecular structure and function, indicating a need of alternative hypotheses. Our analysis revealed that the adaptive embedment of dipeptides and domains in the two waves of recruitment and innovation continues to occur but at different rates. The early dipeptides provided ground materials for combinatorics into domains along the entire protein history. Albeit introduction of relatively newer dipeptides and domains seems ongoing and their continued involvement in recent combination events with older domains suggests that older dipeptides are still evolutionarily active as evolvable forms and not some relics destined for extinction. Thus, the early establishment of highly conserved genetic motifs of dipeptides and their combinatorics defined a hierarchical structural complexity, tightly interrelated throughout protein history. Remarkably, phylogenetic studies demonstrate similar dynamics materializing with domain structures and their combinatorial use in multidomain proteins (Wang and Caetano-Anollés, 2009;

Aziz and Caetano-Anollés, 2021). These alternate views notably falsify evolutionary scenarios of sequential and separate growth of dipeptide and domain repertoires. Although highly conserved evolutionarily, dipeptides are not ‘molecularly canalized’ that resist evolutionary change rather continue to express in different structural contexts. Historically, only functionally useful prior forms prevailed when they provided properties that extend the persistence of the emerging cells, including membrane stability, transport modulation, bioenergetics, and peptide-cofactor biosynthetic functions.

The connectivity of domain and dipeptide components of the D-F network gradually evolved from 3 and 1, respectively, to a global average of 22.36 (± 0.467) (max 84) dipeptides per domain and 81.56 (± 3.838) (max 411) domains per dipeptide, respectively (Supplementary Fig. S7). Remarkably, although starting at 1, dipeptide connectivity gradually modulates up, almost doubling from ~ 3 , to ~ 6 domains per dipeptide during early ~ 1 Gy ($nd < 0.2$) of protein evolution. The domain connectivity first modulates down instead, from 3 to ~ 1.67 (almost half) by first ~ 0.5 Gy ($nd < 0.1$), and then up to ~ 10 (more than triple) dipeptides per domain. Thus, cooption of ancient dipeptides by domains was more robust than enhancements of economy in the number of dipeptides in typical active sites of domain structures. These dual hourglass trends suggest a frustrated dynamics of growth that promise continual innovation.

The GLMs of the degrees of the dipeptide networks, showed that the number of interactions among the dipeptides, which should be interpreted as sharing of proteomes (Supplementary Fig. S8A) or domains (Supplementary Fig. S8B), significantly relate to the dipeptides. The p-value output by the ‘Chi Square’ tests of ANOVA for both domain and proteome-sharing based dipeptides networks was $2.2e-16$ and AIC scores of the main models were 6147 and 6263, respectively, which

were much higher than the scores 6499610 and 25910 of the corresponding null models (Supplementary Fig. S8). These statistical results, plots of residuals and accompanying graphs reject the null hypothesis e) the domain and genome-sharing based dipeptide network degree does not correlate to dipeptide instances.

The residuals and QQ plots (Supplementary Fig. S9) derived from the GLMs of the degrees of dipeptides also intimate that attainment of high degree is characteristic of significantly many dipeptides, which are not monopolized by just a few. The widely spread variance of dipeptides degrees in species (Supplementary Fig. S10), supported by the multi modal distribution of dipeptides abundances (Supplementary Fig. S11), specifically, the dipeptides abundances over the ages of domains in the domain-sharing based dipeptides network (Supplementary Fig. S12A) and the network diameter (Supplementary Fig. S12B) suggests strong richness of scale (Tanaka, 2005) in D-F, a significant network property that we further explore.

4.3.5 The dipeptide network based on domain-sharing is not strictly scale-free

Contrary to scale-free networks (Barabási and Albert, 1999; Jeong *et al.*, 2000) where the network's degree distribution can be modeled as a power-law, owing to few nodes hijacking connectivity, a scale-rich network (Tanaka, 2005) would comprise of many hubs with varying level of connectivity trended towards higher values. It is evident from our network analysis that not only the domain (Supplementary Fig. S5) and proteome-sharing (Supplementary Fig. S10B) based dipeptide networks degrees are multimodal, with medians larger than the means (Supplementary Figs. S7 and S10C), the power law distributions do not fully describe the dipeptides connectivity. The proteome-sharing based dipeptide network does not fit power-law distribution with p-value 4.39×10^{-6} (over a 95% confidence interval), higher KS stat 0.16 and

alpha 57.95, as given by Kolmogorov-Smirnov test (Clauset, Shalizi and Newman, 2009) (Supplementary Fig. S13A). However, domain-sharing shows some signs of scale-freeness with the p-value 0.42, lower KS stat 0.11 and alpha 15.07 (Supplementary Fig. S13B). Still, the results reject the null hypothesis f) domain and genome-sharing based dipeptide networks are scale-free, and in lieu of most biological networks, specifically metabolic networks (Tanaka, 2005), suggest that dipeptide networks ought to be scale-rich and self-dissimilar.

Growing networks that follow the preferential attachment principle have large, highly connected hubs that attract increasingly more links in a ‘rich-get-richer’ fashion. In these highly inhomogeneous networks, which are ubiquitous in biology, the probability $P(k)$ of a node linked to k nodes (i.e., the fraction of nodes with k links or k -neighbors) decays as a power law, $P(k) \sim k^{-\gamma}$, devoid of a characteristic scale. ‘Scale-free’ networks in general have regression exponents $\gamma = 2.1$ – 2.4 , with a typical heavy-tailed distribution in which very few nodes have high connectivity degrees (Strogatz, 2001). For metabolic reaction networks, $\gamma = 2.2$ in all organisms (e.g., (Jeong *et al.*, 2000)). To test the evolving D-F network and its projections for a tendency to follow the scale-free properties in their degree distributions, we observed the chronological cumulative connectivity (represented by links or arcs) of the growing networks and tested them for power law behavior with appropriate statistics (Fig. 4.6A; Supplementary Fig. S14). Remarkably, we noticed a power law tendency as an emergent property, along with associated generative models, in the ancient portion of the D-F network, but not significantly in its projections, especially during earlier part of the timeline.

Although the domain component began with somewhat “scale-rich” (Tanaka, 2005) tendencies, a few statistics failed to reject power law properties in both the most ancient dipeptide and domain

components until $nd \sim 0.053$. The domains then continue to mostly connect in a scale-free fashion until $nd \sim 0.265$ and $nd \sim 0.400$, with major interruptions to power law behavior from $nd \sim 0.127$ – ~ 0.176 and $nd \sim 0.269$ – ~ 0.363 (Fig. 4.6A; Supplementary Fig. S14). The dipeptide component remains self-dissimilar (scale-rich) from $nd \sim 0.057$ to $nd \sim 0.249$ and shows self-organization (scale-freeness) until $nd \sim 0.441$ and $nd \sim 0.539$, with an interruption to power law behavior from $nd \sim 0.445$ – ~ 0.510 (Fig. 4.6A; Supplementary Fig. S14). The more recent connectivity of both domains (since $nd \sim 0.404$) and dipeptides (since $nd \sim 0.543$) of the D-F bipartite network continues to be “scale-rich” until modern times. Thus, scale-free properties were found evident in co-options of i) earlier domains and dipeptides, ii) older domains by multiple dipeptides, and iii) relatively younger dipeptides by multiple domains (Fig. 4.6A) (with short time periods of exceptions in ii and iii). The most prominent indicator of failure to reject was the Kolmogorov-Smirnov (KS) statistical test of power law fit (Newman, 2005; Clauset, Shalizi and Newman, 2009). High p-values of the KS test (≥ 0.1) and relatively lower values of the KS fit statistic (< 0.5) fail to reject network data being drawn from the fitted power-law distribution. Similarly, the exponent of the fitted power law distribution (α), which is $\alpha > 2.5$ when the assumption of probability of power law fit $P(X = x^{-\alpha})$ is met in this case, was incidentally lower, for example, for later degree distributions of domain and dipeptide component of the growing D-F network. These crisscross patterns are indicative of a biphasic rise and fall of power-law property. The log-likelihoods of the fitted power law parameters gradually deviate towards values lower than zero for most network events of the timeline, showing typical power law distributions less probable, in general.

The statistical analyses of the timeline of growing D-F networks therefore confirms a prevalent property of power law emergence and transference. While the earlier dipeptide component of the bipartite network shared power law behavior with the domain component, these ancient

components went through two cycles of an exchange of scale-free properties, from dipeptide to domain and then from domain to dipeptide, as molecular functions developed in protein evolution (Fig. 4.6A; Supplementary Fig. S14). *Log*-linear regression models overlapping the power law curves showed that the coefficient of power law decay γ for the domain portion of the D-F network remains mostly linear or sublinear until $nd \sim 0.041$ ($\gamma \sim 0.4 \sim 1.0$) and since $nd \sim 0.086$ ($\gamma \sim 0.5 \sim 1.0$), with a short duration of moderate power law scale ($\gamma \sim 1.1 \sim 1.5$) from $nd \sim 0.045 \sim 0.078$, $\gamma = 0.839 \pm 0.009$ on average and $\gamma = 1.481$ maximum. For the dipeptide portion, however, γ starts with a moderate power law scale ($\gamma \sim 1.6$), $nd \sim 0.01$), followed by a short duration of mostly linear or sublinear scale ($\gamma \sim 0.1 \sim 1.0$) from $nd \sim 0.020 \sim 0.069$, but then shifting to a moderate to moderately high power law scale ($\gamma \sim 1.1 \sim 1.9$) from $nd \sim 0.073 \sim 0.384$, and finally reverting back to a linear or sublinear scale ($\gamma \sim 0.3 \sim 1.0$) for the rest of the timeline, in a somewhat antisymmetric pattern to that for domain, with an average $\gamma = 0.791 \pm 0.032$ and max $\gamma = 1.929$. The coefficient of determination (R^2) of $\sim 90\%$ supports the linear models for both the domain and dipeptide components. Thus, the extent of preferential attachment in dipeptide evolution is less prominent than that reported for metabolic networks (Ravasz *et al.*, 2002). Noticeably, dipeptide ($\gamma = 0.078 \pm 0.0140$) and domain ($\gamma = 0.132 \pm 0.0096$) network projections maintained only a sublinear scale. The extraordinary dual yin-and-yang like power law transference between dipeptide and domain components is a reconfirmation of a continuing global scaling phenomenon of a biphasic emergence of biological modules (Mittenthal, Caetano-Anollés and Caetano-Anollés, 2012), which we elaborate further. Note: the power law arbitrary scales used above are defined as follows: linear or sublinear: $\gamma \leq 1.0$, moderate: $1.0 < \gamma \leq 1.5$, moderately high: $1.5 < \gamma < 2.0$.

4.3.6 Hierarchical modularity at the dipeptide level of protein makeup

Modular networks have most of their node connectivity within bounds of communities (modules) (Newman and Girvan, 2004). Modularity reciprocates the scale-free property of biological networks by equalizing the degree distribution of nodes in communities (Jeong *et al.*, 2000; Overbeek *et al.*, 2000; Wagner and Fell, 2001). However, these opposing trends reconcile when modules are integrated hierarchically (Ravasz *et al.*, 2002). A primary measurement of modularity is the *average clustering coefficient* (C), a ratio of triangles (graph cycles of length 3) to the connected triads in the graph, averaged over all nodes, while ignoring the direction and weights of the edges (Wasserman and Faust, 1994; Ravasz *et al.*, 2002; Barrat *et al.*, 2004). Since a bipartite network cannot have a meaningful C due to absence of triangles, modular organization of D-F was investigated through its projections (Fig. 4.6B; Supplementary Fig. S15). The domain and dipeptide networks exhibit C values of $\sim 0.816 (\pm 0.0062)$ and $\sim 0.824 (\pm 0.0069)$, respectively, significantly higher than ~ 0.6 reported for metabolic networks (Watts and Strogatz, 1998; Wagner and Fell, 2001; Ravasz *et al.*, 2002). High C of D-F network projections suggests that densely intra-connected modules of dipeptide motifs and domain structures are integrated by few inter-modular links. The D-F network thus embodies a highly consolidated modular structure.

C sharply declines with network size N for a scale-free model (Albert and Barabási, 2002), as $N^{-0.75}$, while highly modular networks are independent of N (Ravasz *et al.*, 2002). For the domain and dipeptide networks, C regressed positively with N as $N^{+0.000078}$ and $N^{+0.000580}$, and with age nd of the networks as $nd^{+0.18}$ and $nd^{+0.28}$ (Fig. 4.6B; Supplementary Fig. S15), respectively, confirming the modular structure of the evolving networks. As expected, C of strictly power-law (Barabási) reference controls were zero (Newman, Strogatz and Watts, 2001). The timeline of growing

domain and dipeptide networks revealed anticorrelated trends of modularity and scale-free properties. For example, the C of domain network rises from ~ 0.6 at $nd \sim 0.028$ to ~ 0.7 at $nd \sim 0.041$ and then hits a plateau of ~ 0.8 at $nd \sim 0.180$, while that of dipeptide network initially fluctuates between ~ 1.0 and ~ 0.8 from $nd \sim 0.000$ to ~ 0.020 to ~ 0.028 to ~ 0.045 , but then stabilizes to ~ 0.6 at $nd \sim 0.057$ before gradually rising to ~ 0.7 at $nd \sim 0.126$ and then to ~ 0.8 at $nd \sim 0.249$, finally hitting the plateau of ~ 0.9 at $nd \sim 0.580$. These patterns are reciprocal to the power law trends. In domain network, the KS fit indegree statistic remains lower in a ~ 0.1 – ~ 0.4 range from $nd \sim 0.020$ – ~ 0.078 , indicating preferential attachment, but then has a cycle of rising to ~ 0.5 , ~ 0.6 , ~ 0.7 and ~ 0.8 at $nd \sim 0.078$, ~ 0.155 , ~ 0.171 and 0.196 , and then again to ~ 0.5 , ~ 0.6 , and ~ 0.7 at $nd \sim 0.249$, ~ 0.359 , and ~ 0.559 , respectively, rejecting power law behavior, before dropping back to ~ 0.6 at $nd \sim 0.824$ and then to ~ 0.4 at $nd \sim 0.837$, indicating scale-free behavior in modern domain networks. There were, however, some characteristic dips in KS fit to ~ 0.3 at $nd \sim 0.135$ and to ~ 0.4 at $nd \sim 0.767$ in the domain network, indicating brief spasms of preferential attachment at these transitory periods of recruitment. (Fig. 4.6B; Supplementary Figs. S14 and S15). In dipeptide network, KS fit remains in ~ 0.1 – ~ 0.5 range until $nd \sim 0.094$, and then in a very high ~ 0.6 – ~ 0.9 range throughout the timeline, except a characteristic dip to ~ 0.5 at $nd \sim 0.653$ and to the same value of ~ 0.5 during $nd \sim 0.706$ – ~ 0.845 (Fig. 4.6B; Supplementary Figs. S14 and S15). This indicates that dipeptide network is more highly modular than domain network and is less influenced by the scale-free properties, except of course in the initial phases and at the transitory stages at $nd \sim 0.653$ and $nd \sim 0.706$. These counteracting trends of modularity that follow those of preferential attachment with a lag of $\Delta nd \sim 0.020$ in the domain and dipeptide networks are reflective of an emergent power law behavior of a modular D-F bipartite network. Our findings thus highlight a transfer event of scale-free properties from dipeptide to domain components of the dipeptide functionome,

involving a generative cycle of modular and hierarchical structure of interacting dipeptide motifs and domain structures.

We verify this conjecture with an accompanied analysis of the dynamics of three measures of network modularity along the timeline of growing networks. The Newman-Girvan (*NG*) algorithm iteratively calculates edge ‘betweenness’, i.e., the maximum number of shortest paths running through an edge, while systematically removing edges scoring high in inter-community centrality (Newman and Girvan, 2004). This perpetual removal uncovers the modular structure of a network measured by the *NG* index. *NG* partitioned by age (NG_{age}) ranges from -1 to 1 , with positive values indicating tendency to connect to modules of the same age, while negative values indicating tendency across. *NG* partitioned by VOS (NG_{vos}) describes the cohesiveness of VOS memberships (Van Eck and Waltman, 2007; Waltman, van Eck and Noyons, 2010). The Fast Greedy Community (*FGC*) detection algorithm provides a hierarchical perspective of agglomerative community structure (Clauset, Newman and Moore, 2004). Noticeably, NG_{vos} and *FGC* indices produced mirrored patterns of growth of modular cohesiveness and agglomerative structure by age for all growing networks. Conversely, NG_{age} indicated divergent patterns of age-associativity among the bipartite D-F network and its domain and dipeptide network projections. The D-F and dipeptide networks highly showed age-associative origins with $NG_{age} \sim 0.2 \sim 0.3$ until $nd \sim 0.184$ and $nd \sim 0.041$, respectively, and ~ 0.1 until $nd \sim 0.339$ and $nd \sim 0.163$ (more sporadically in dipeptide with $NG_{age} \sim 0.0$ sometimes), respectively. After those time events, an $NG_{age} \sim 0.0$ during rest of the timeline indicated “carefree” association to age, supporting uniform cooption and recruitment age-wise patterns, in the D-F and dipeptide networks timeline (Fig. 4.6B; Supplementary Fig. S15). The domain network, however, shows strong age-dissociative nature in the beginning with NG_{age} of ~ -0.5 , -0.2 and -0.1 until $nd \sim 0.012$, ~ 0.041 and ~ 0.073 , respectively, before stabilizing to an

age-wise “indifferent” cooption pattern at $nd \sim 0.078$ with $NG_{age} \sim -0.0$ for the rest of the timeline (Fig. 4.6B; Supplementary Fig. S15). These network modularity growth patterns are indicative of more robust age-wise cohesive recruitment in the dipeptide component than domain structures, with dipeptide motifs being tightly coupled by age during dipeptide functionome emergence. This trend of course plateaued to an equilibrium as network agglomerative modularity matured and existing forms engaged in widespread cooption of emerging constructs throughout the timeline. This recruitment trend is also demonstrated in the pairwise NG_{age} heat maps of the D-F network by a red sigmoidal signal during earlier part of the timeline (first three panels) diffusing into a pervasive (red pixelated) pattern in modern networks (last two panels) (Fig. 4.6C; Supplementary Video 3). These modular matrix representations combined with the analytics from the power law and modularity statistical indicators reflect a parallel development of hierarchy that encouraged genetic modules to cluster within modules in a fashion not distinct from the clustered scale-free organization of metabolic networks (Ravasz *et al.*, 2002) (Supplementary Fig. S16; Supplementary Video 4). Nevertheless, the growth of dipeptide and domain networks while initially driven opposing age-associative and age-dissociative recruitment patterns, respectively, was quickly but moderately counteracted by age-bound modularity. Thus, the transfer of scale-free properties between dipeptide motifs and domain structures involves a hidden switch to a hierarchically modular structure, which occurred ~ 3.4 Gy ago. The timing of this switch has been previously discovered to overlap with the early development of genetic code specificity in emerging aminoacyl-tRNA synthetases and the ribosome (Caetano-Anollés, Wang and Caetano-Anollés, 2013; Aziz, Caetano-Anollés and Caetano-Anollés, 2016) that was moderated by the OB-fold structure.

4.4 Conclusions and Future Directions

Tracing phylogenetic age onto evolutionary film of dipeptide-footprint networks and their projections verified the two waves of structural and functional innovation involving founder ‘*p*-loop’ and ‘winged helix’ domain structures (Aziz, Caetano-Anollés and Caetano-Anollés, 2016; Aziz and Caetano-Anollés, 2021). We found that the growth of recruitment of dipeptide units into domain structures is modern and highly modular. The records of molecular footprints featured hierarchical modularity and a transformational power law behavior. The D-F network showcases communities of interacting structural parts defining modular molecular structural classes. Our analyses verifies that such hierarchical modularity evolved through a biphasic process of diversification (Mittenthal, Caetano-Anollés and Caetano-Anollés, 2012). Biphasic patterns are prevalent in the size, dipeptide makeup, and loop-mediated flexibility of proteins (Caetano-Anollés, Wang and Caetano-Anollés, 2013; Debès *et al.*, 2013), and possibly linked to their intrinsic disorder. This adaptation to a biphasic pattern of change favors to be a general phenomenon in network biology.

The data generated in the present study will be used in the future to analyze evolving timed networks of dipeptides based on the sharing of both domains and proteomes. This merits further investigation into temporal behaviors of randomness in network structure and its correlation with scale-freeness and modularity aspects explored here.

4.5 Methods

4.5.1 The Fold Families (domains)

The 1,475 fold families selected for this study (Caetano-Anollés, Wang and Caetano-Anollés, 2013) were represented as concise classification strings, defined at “Fold Family” level of

Structural Classification of Proteins (SCOP) (Murzin *et al.*, 1995) hierarchy. For example, the catalytic domain of tyrosyl-tRNA synthetase [EC 6.1.1.1] corresponds to the c.26.1.1 domain, in which c represents the protein class (a/b proteins), 26 the fold (adenine nucleotide alpha hydrolase-like fold), 1 the fold superfamily (nucleotidyl transferase superfamily), and 1 the fold family (Class I aaRSs, catalytic fold family).

4.5.2 Calculation of the Ages of domains

The age or *nd* attribute values of domains were derived directly from the rooted phylogenetic tree of protein domain architectures (Caetano-Anollés, Wang and Caetano-Anollés, 2013). A Python script was used to count the number of nodes from the root (base) of the tree to each leaf node and provide the node distances in a relative zero-to-one scale (Caetano-Anollés, Wang and Caetano-Anollés, 2013). A fundamental feature of the trees of domain structures fostering the derivation of relevant ages is a high imbalance (Wang and Caetano-Anollés, 2009). The tree imbalance was resultant from the accumulation of domain structures in proteins and proteomes (Wang *et al.*, 2011), instead of mere node density, and thus represents a true evolutionary process.

4.5.3 The Species

An eclectic sample of 1,561 species was selected for the original study from the three super kingdoms: Archaea, Bacteria and Eukarya (Caetano-Anollés, Wang and Caetano-Anollés, 2013), given that their protein sequences were available in the Protein Data Bank (PDB) (June 20, 2012). Each species was assigned an acronym. The label for species with acronym “dM” and proteome size 3,651 was unavailable.

4.5.4 The Proteome Size

Number of proteins or proteome size of each of the 1,561 species (Caetano-Anollés, Wang and Caetano-Anollés, 2013) was recorded from the PDB (June 20, 2012).

4.5.5 The Dipeptide (dipeptide) Abundances and Ages

Abundance values of the dipeptides in the secondary structures of species were calculated using a non-culled and a culled set of protein entries from the PDB (Caetano-Anollés, Wang and Caetano-Anollés, 2013). Secondary structures were assigned using the Dictionary of Secondary Structures of Proteins (DSSP) (Kabsch and Sander, 1983). The non-culled set including 204,531 domain sequences (51,392,487 amino acids) was directly downloaded from the PDB (June 20, 2012). The culled set was selected using the protein sequence-culling server PISCES (Wang and Dunbrack Jr, 2003).

For calculation of dipeptide frequencies in domains (Caetano-Anollés, Wang and Caetano-Anollés, 2013), SCOP (Murzin *et al.*, 1995) structures were assigned to PDB sequences with the Hidden Markov Models (HMMs) from SUPERFAMILY (Gough *et al.*, 2001). The PDB sequences were screened for $\leq 25\%$ identity using PISCES (Wang and Dunbrack Jr, 2003). PDB sequences with two or more domains or unassigned ranges longer than 30 amino acids were not included. The final PDB sequence set comprised of 2,384 sequences covered by 1,475 domains.

For calculation of dipeptide ages, since a dipeptide is embedded in an domain structure and both dipeptides and domains describe functional and structural abstractions, the age of domains can be directly transferred to dipeptides. We considered two likely schemes to do this: (i) the age of an dipeptide is the age of the most ancient associated domain, or (ii) the age of an dipeptide is the age

of the more recent from a pair of most ancient associated domains (or the age of a singly associated domain). These schemes consider the age of an dipeptide as either the age of the first structural scaffold or the age when the dipeptide function is first transferred between structural scaffolds, respectively. While these both provide similar age mappings, we only derive mappings using the second more conservative scheme.

4.5.6 Data and Project Design

The data sample comprised of 1,561 chosen species from the three super kingdoms: Archaea, Bacteria and Eukarya (Caetano-Anollés, Wang and Caetano-Anollés, 2013), based on availability of protein sequences in the PDB (June 20, 2012). Furthermore, the fold families were sampled from a set of 2,384 PDB sequences, covered by 1,475 domains. In the abundance data, 400 permutations corresponded to each dipeptide frequencies, from AA to VV (Supplementary Fig. S11). In the dipeptide network's degree data, each row corresponded to connectivity values of one of the 400 dipeptides.

For multivariate and principal component analyses, the respective data of dipeptide abundances for species and fold families were organized in a matrix form, with each column corresponding to one of the 400 dipeptides, from AA to VV (Supplementary Fig. S11). For running the generalized linear models, dipeptide abundance data was melted or factored (Supplementary Fig. S12) using R's (R Core Team, 2014) "reshape2" package in order to represent the abundances of dipeptides as factorial categories (Supplementary Fig. S11). The vectors of abundances were then respectively modeled against the main effects of proteome size (Supplementary Fig. S11) and age (Supplementary Fig. S12). In network analysis the domains data set was pre-processed through R scripts and Microsoft Excel spreadsheets. Pajek (Mrvar and Batagelj, 2016) and "igraph" package

of R (R Core Team, 2014) were used for analysis and visualization of networks (Supplementary Fig. S12), in particular, for the calculation of dipeptides network's degree (Supplementary Fig. S10). The dipeptide degree values were also used in computation of a GLM using Poisson family (Supplementary Fig. S4) and testing for preferential attachment principle (Barabási and Albert, 1999; Jeong *et al.*, 2000) using Kolmogorov-Smirnov test (Clauset, Shalizi and Newman, 2009) (Supplementary Fig. S13).

4.5.7 Data Summaries

Data summaries for species and fold families-based dipeptide abundances were computed in R (R Core Team, 2014), respectively (Supplementary Fig. S11).

4.5.8 MANOVAs

dipeptide abundances in a form of 400 column matrix was input to multifactorial MANOVA for species data set, having main effects of 'Number of Dimers' along with proteome size (Supplementary Fig. S1A). Both the effects and the interaction were found significant. MANOVA on main effect of ND_val or Age in domains data set was also found significant (Supplementary Fig. S1B).

4.5.9 PCAs

PCs for the standardized abundance matrix of 400 dipeptides were computed from the species (Supplementary Fig. S2A) and domains (Supplementary Fig. S2B) data sets. Standardization was accomplished by calculating z-scores with mean equal to zero and variance one. The only PCs kept were with standard deviation greater than one. Scree plots presented the perspective of

retention of PCs based on the elbow or inflection point in the Inertia (Supplementary Fig. S2C, D). Pair of PCs explaining highest variances were plotted (Supplementary Fig. S3) to identify classes.

4.5.10 Melting or Factoring Data

In order to run the GLMs, the dipeptide abundance data in matrix form was melted to co-align the abundance values as factorial categories of dipeptides along the main effects of proteome size in species (Supplementary Fig. S2A) and age in domains (Supplementary Fig. S2B) data sets, respectively.

4.5.11 Abundance Histograms

Histograms of melted data demonstrated that the dipeptide abundances for both the species (Supplementary Fig. S11E) and domains (Supplementary Fig. S11F) were rightly skewed.

4.5.12 Box Plots

The Box and Whisker's Plots of the melted data sets for the species (Supplementary Fig. S11A) as well as domains (Supplementary Fig. S11B) showed high variance in abundances of the 400 dipeptides.

4.5.13 Scatter or XY Plots

The scatter plot for the melted species data was generated among the dipeptide abundances and the proteome size or number of proteins (Supplementary Fig. S11C). With a few vacancies in the discrete values of proteome size, the overall trend of dipeptide abundances was found to linearly increase with the proteome size. In the melted domains data set, the XY-plot of dipeptide

abundances against ND_val or age (Supplementary Fig. S11D) confirmed that, with just a few bins missing from the continuous age values, overall distribution of dipeptide abundances across ages was multi modal.

4.5.14 Age Wise Scatter Plot

The ND_val or age of domains posed an opportunity to plot the dipeptide abundances grouped as age factor (Supplementary Fig. S12). It was observed that the mountain ranges of dipeptide abundances were typically higher for post middle age domains, with an overall multi modal pattern.

4.5.15 GLMs

Since dipeptide abundances or frequencies data essentially resembles count data, ‘Poisson’ glm family was chosen to run the GLMs. The abundances were modeled as response variables against the respective main effects of proteome size and age, in each of species (Supplementary Fig. S4A) and domains (Supplementary Fig. S11B) melted data. The main models with one main effect were compared with the null model using ‘Chi Square’ test of ANOVA. AIC scores for the main models were also computed (Supplementary Fig. S8). In both data sets, abundances were found to be significantly described by the main effects. In domains melted data (Supplementary Fig. S8B), the normalized dipeptides abundances were transformed to integers by multiplication with 10×10^9 for usage compatibility with ‘Poisson’ family.

4.5.16 Network Generation, Visualization and Analysis

Given the normalized dipeptide abundances in domains and species, the melted or factored form of the data helped create networks of dipeptides. Using a sparse matrix criterion of connectivity, the dipeptides with non-zero abundances in domains or with abundances larger than average in species were chosen to be connected to the corresponding entity. Hence, the sparse matrix criteria led initially to bipartite networks among domains and dipeptides (Supplementary Fig. S12B), and proteomes and dipeptides (Supplementary Fig. S10), comprising of 230,833 links extracted from 590,000 records; and 150,916 links extracted from 624,400 records, respectively. Subsequently, Pajek (Mrvar and Batagelj, 2016) was used to extract single mode networks of dipeptides only, based on the sharing of domains (Supplementary Fig. S12) or proteomes (Supplementary Fig. S10), by the dipeptides in the bipartite networks. Pajek (Mrvar and Batagelj, 2016) enabled to calculate the degree of the dipeptide networks. The dipeptides network degree data was purged to circumvent the ‘NA’ value problem because of the presence of the dipeptide “NA”.

Networks were visualized and analyzed using Pajek (Mrvar and Batagelj, 2016) and R’s *igraph* package (Csardi and Nepusz, 2006). Community-based layouts of the networks were generated using the Visualization of Similarity (VOS) clustering method (Van Eck and Waltman, 2007; Waltman, van Eck and Noyons, 2010). Network properties were analyzed with code constructed using graphing packages and tools of R (Ihaka and Gentleman, 1996; R Core Team, 2014). A detailed description of data files, partitions and functions used to analyze network data, produce charts and plots, compute power law statistics and modularity indices, and construct waterfall diagrams can be found in the Supplementary Materials and Methods.

4.5.17 Networks Degree Histograms

The histograms of the degrees of species' genome-sharing based dipeptide networks with 400 breaks (inclusion of all dipeptides) show a multi modal distribution (Supplementary Fig. S10B), which is slightly rightly skewed.

4.5.18 Networks Degree Box Plots

The box plots of the degrees of species' genome-sharing based dipeptide networks indicate high variance in degree values, with median slightly above the mean (Supplementary Fig. S10C), with a few outliers.

4.5.19 Networks Degree Scatter Plots

The scatter plots of the degrees of domain and species' genome-sharing based dipeptide networks against dipeptides show a significant level of variance, with wide spread of degree data points for dipeptides (Supplementary Figs. S10A and S12A). However, the latter network shows degree data points restricted in the third quartile.

4.5.20 GLM of Networks Degree

Generalized linear model of the degrees of domain and species' genome-sharing based dipeptide networks was computed using 'Poisson' glm family because degrees are essentially counts. The degrees were modeled as response variable against the main effect of dipeptides (Supplementary Fig. S4). The main models with one main effect were compared with the null models using 'Chi Square' test of ANOVA. AIC scores of the main models were also computed and compared with

those of the null models (Supplementary Fig. S4). It was found that the dipeptide degrees are significantly described by the dipeptides.

The residuals of the glm of the networks' degrees were also plotted (Supplementary Fig. S9). The density of predicted values from middle to upper end shows that higher degree values span more than few dipeptides. Observation of a couple of outliers from QQ plot towards the negative end bolsters the idea of scattered high degree values, with median slightly higher than the mean (Supplementary Fig. S9).

4.5.21 Power Law Fit of Network's Degree

The degree or level of connectivity of the domain and species' genome-sharing based dipeptide networks was each put to test for scale free network behavior using $P(k)$ vs. k (probability of having k -neighbors vs. k) and $\log\text{-}\log$ ($\log$ of $P(k)$ vs. $\log$ of k) mathematical curves, with linear regression models to derive γ of the power law and the determination coefficient (R^2). γ is the absolute slope of the log linear model. Higher γ indicates increased tendency towards preferential attachment. R^2 describes the percentage of the data fitting the linear model. When both γ and R^2 are high, scale free behavior should be considered strongly supported. Other power law statistics included: (i) KS fit statistic, which compares the fitted distribution with the input degree vector; (ii) the KS p -value, with the null hypothesis of data being drawn from the power law distribution (Newman, 2005; Clauset, Shalizi and Newman, 2009); and (iii) the exponent of the fitted power law distribution (α), which assumes $P(X=x)$ is proportional to $x^{-\alpha}$. Lower KS fit score, larger KS p -value (≥ 0.05), and higher α suggest better fit to power law distribution. We also determined the maximum log likelihood of the fitted parameters. Reference networks were created using 'Barabási' (Barabási

and Albert, 1999) methods of R's *igraph* package (Csardi and Nepusz, 2006) to simulate power law and extended age-dependent control models for the corresponding networks.

4.5.22 Network Modularity

We studied modularity with six indices: (i) The VOS Quality index (VQ), was generated by the Pajek layout algorithm that considers weights of links (edges/arcs) as similarities. Communities were iteratively drawn closer based on similarity and the quality index of the final layout with least crossings and closest clusters was given. VQ is then calculated as $\sum_{i=1}^c \sum_{j=i+1}^c (e_{ij} - a_i^2)$, where c is the number of communities. e_{ij} is the fraction of edges with one node v in community i and the other w in community j , given as $\sum_{vw} (A_{vw}/2m)$ with $1_v \in c_i$, $1_w \in c_j$, where m is the sum of weights in the graph and A_{vw} = the weighted value or 0, indicating presence or absence of edge between the nodes v and w , respectively, in the adjacency matrix A of the network. Finally, a_i is the fraction of weighted k neighbors that are attached to the nodes of a community i , i.e $k_i/2m$ (Van Eck and Waltman, 2007; Waltman, van Eck and Noyons, 2010); (ii) The Clustering Ratio (C-ratio) is the ratio of the number of clusters to the size of an inter connected node set; (iii) The average Clustering Coefficient (C) describes the mean ratio of triangles to connected triads over all nodes in the simplified (undirected/unweighted) network (Wasserman and Faust, 1994; Ravasz *et al.*, 2002; Barrat *et al.*, 2004). C is meaningful only for unimodal graphs (Newman, Strogatz and Watts, 2001). We also report coefficients of linear regression over C for domain and dipeptide network projections; (iv) The Fast Greedy Community (FGC) hierarchical agglomeration algorithm detects community structure with linear run time $O(m d \log n) \sim O(n \log^2 n)$, of a network with m edges, n nodes, and depth d of the dendrogram describing its community structure (Clauset, Newman and Moore, 2004); and (v and vi) The *Newman-Girvan* algorithm index (NG), computed with partitions

defined by age (NG_{age}) and VOS clustering (NG_{vos}). NG calculates the modularity of a network based on some classification (partition) to measure how good that classification is in dividing the various node types, indicated by assortative (positive) or disassortative (negative) mixing across modules (Newman and Girvan, 2004). NG equals $1/(2m)\sum_{ij}(A_{ij}-1/(2m)k_i k_j * \Delta(c_i, c_j))$, where m is the collective weights in the graph, A_{ij} are weighted entries in the adjacency matrix, k_i , k_j and c_i , c_j are the weighted degrees and the components (numeric partitions), respectively, of nodes i and j each, and finally, $\Delta(x,y)$ is 1 if $x=y$ and 0 otherwise (Newman and Girvan, 2004). VQ , C -ratio, C and FGC each range from 0 to 1, while the NG indices range from -1 to 1. Higher indices represent strong network modularity at an event. Heatmaps were generated using customized scaled modularity matrices with elements given as $(A_{ij}-k_i k_j/(2m))M_{nd}$, where A_{ij} , k_i , k_j and m are as defined for NG (Newman and Girvan, 2004), and M_{nd} is a network's modularity index at event nd . Dendrograms were calculated as squared Euclidean distance matrices indicating dissimilarities between the cluster means (Borg and Groenen, 2003). The distance (or dissimilarity) matrices were hierarchically clustered with the Ward's minimum variance method aiming at finding compact, spherical clusters (Murtagh and Legendre, 2014).

4.6 Figures

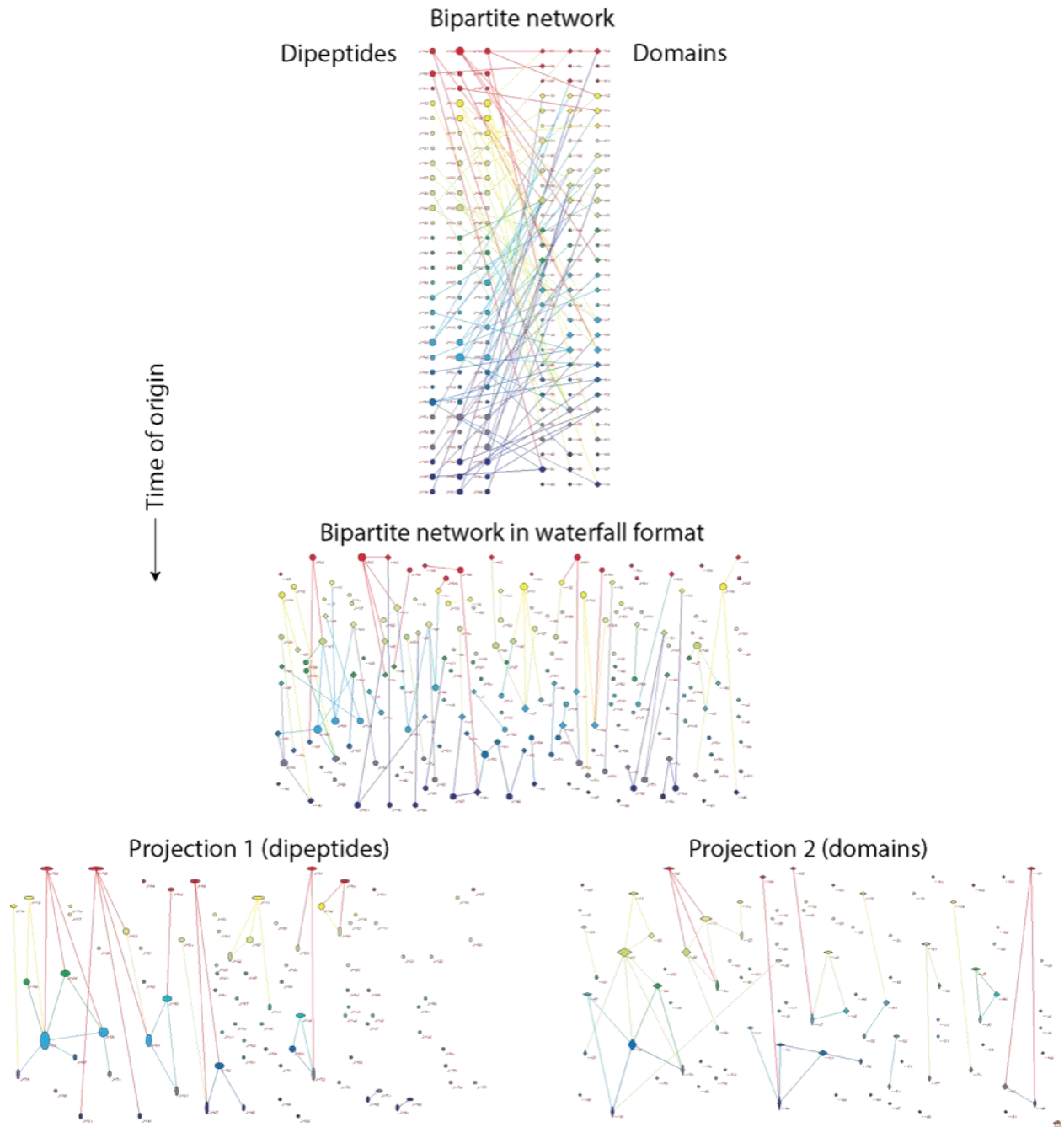


Figure 4.1. Various network forms, features and layouts illustrated through a model miniature network. The diagrams illustrate an undirected bipartite miniature network in bipartite (A) and waterfall (B) layouts its two network projections in waterfall layouts (C,D). The original form has 90 nodes per bimodal partition, 3 nodes per age group, and 9 nodes per age-bin. Nodes are described as symbols, with size proportional to the number of links they establish. The connectivity in the undirected bimodal network was generated at random, strictly between the two partitions (DPs

Figure 4.1 (cont.)

or dipeptides and FFs or domains). The connectivity in the projections was generated based on the sharing instances within partitions and describes construction of directed ‘discrete event’ networks in waterfall format. As time progresses from top to bottom in a waterfall layout, events pan out as the progressive appearance of nodes and links, clustered from left to right. Network growth was made explicit by coloring of nodes according to age (red-to-blue) and that of links by node of origin, by using time-induced arrows or arcs (edges in case of undirected) with density proportional to node connectivity, and by maintaining horizontal and vertical sizes of symbols proportional to the respective alldegree, outdegree and indegree of the nodes. As the network grows with time, the transition from wide to tall symbols facilitates the visualization of the source-sink origination dynamics of recruitments in the projections, with chronological accumulation of connections increasing outdegree of an earlier originating node and widening the node-symbol horizontally, while increasing the indegree of a later receiving node and lengthening the vertical scale of the node-symbol.

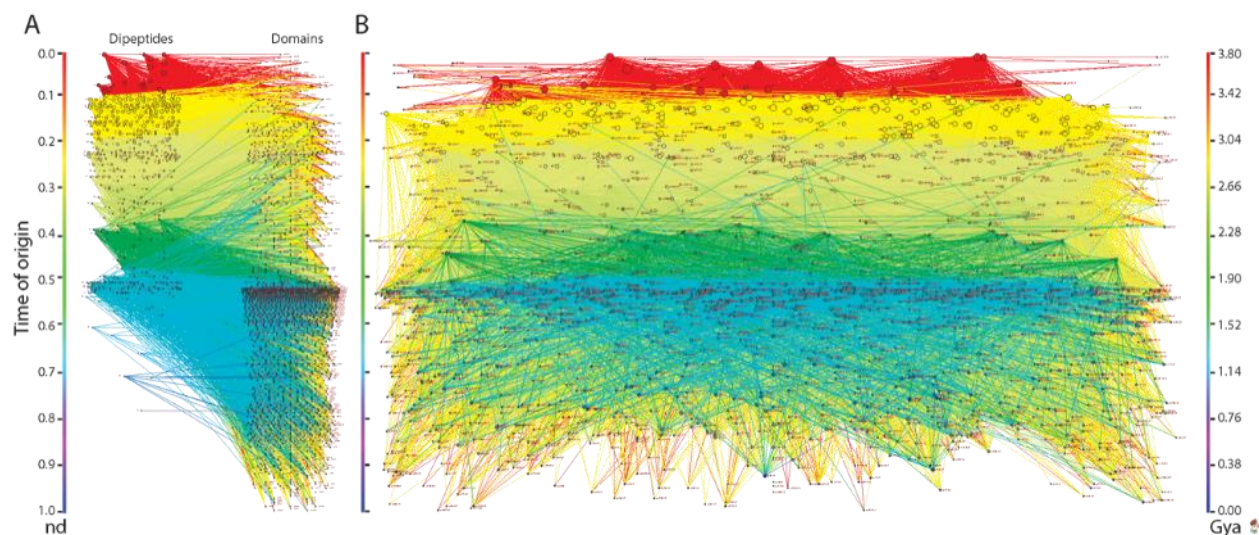


Figure 4.2. The D-F network in bipartite (A) and waterfall (B) layouts. The dipeptide and domain nodes were arranged top-down according to age (*nd*) displayed in a relative 0-to-1 scale, labeled using established SCOP (Murzin *et al.*, 1995) nomenclature, respectively, and colored according to time events (left). Links (undirected) were colored by age of a dipeptide origin node. Ages were also time-calibrated with a molecular clock of domains that spans 3.8 billion years (Gy) of history using fossils and microfossils, geochemical, biochemical, and biomarker data (Wang *et al.*, 2011) (right). The nodes were scaled proportional to their weighted degree, i.e. the sum of the weights of all edges of the nodes. Dipeptide abundances in structural domains in proteomes were not used to weight edges to avoid complication in interpretation of weighted network projections. Red arrowheads indicate the origin of major waves of recruitment in the time event waterfall. The horizontal expansion is dictated by VOS clustering, which elicits formation of modules along the evolutionary timeline (see methods).

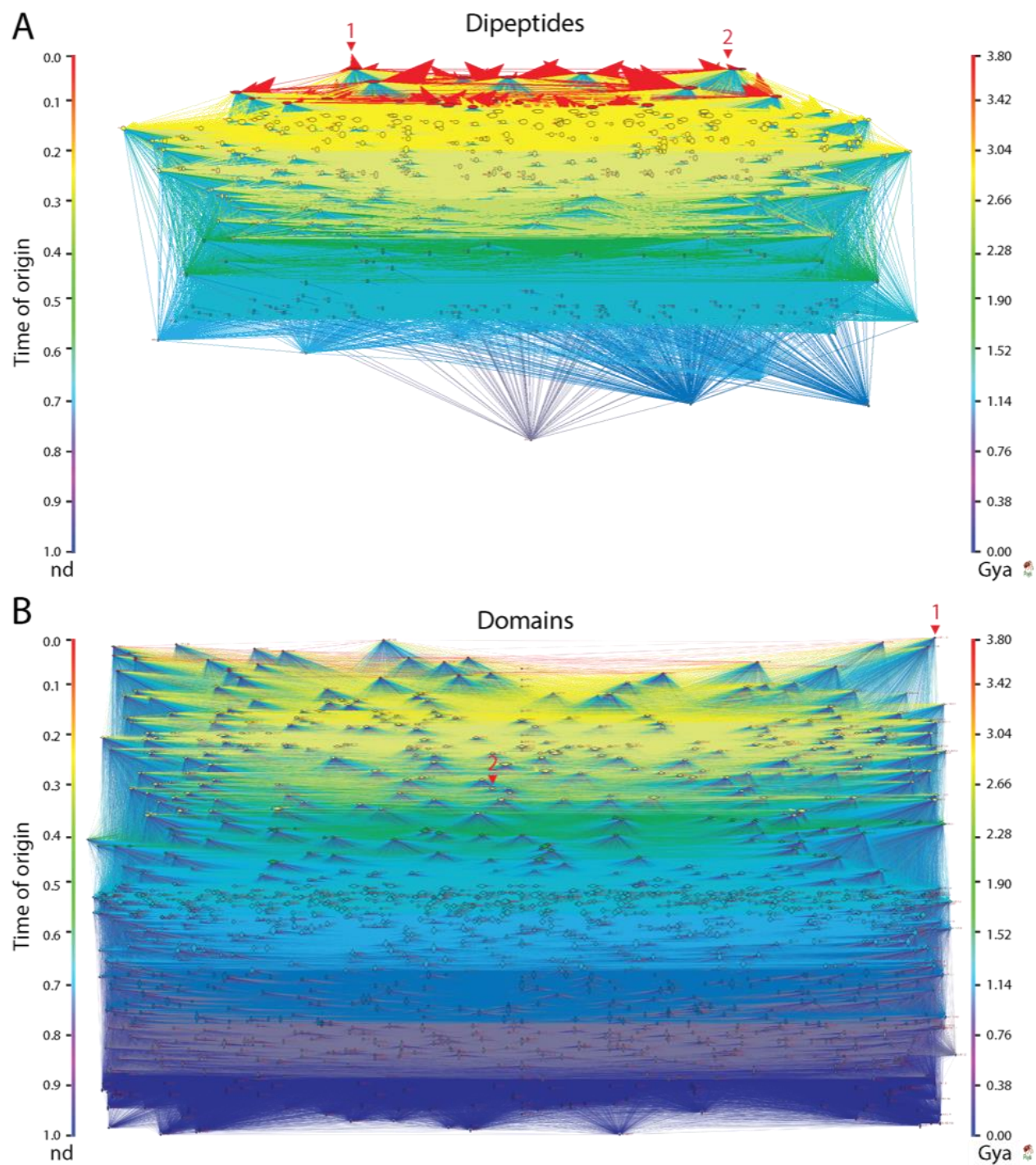


Figure 4.3. D-F network projections in waterfall layout. (A) dipeptide network defined by 400 dipeptides (ellipsoids) and arc connections (arrows) representing sharing of domain structures. **(B)** domain network defined by 1475 domains (rhomboids) and arc connections representing sharing of dipeptide units. The dipeptide and domain nodes in their one-mode graph representations were arranged top-down according to age (*nd*) displayed in a relative 0-to-1 scale, labeled using established SCOP nomenclature (Murzin *et al.*, 1995) and colored according to time events (right). Links (directed) were colored by the age of a destination node. The 2D scale of nodes was kept proportional

Figure 4.3 (cont.)

to their weighted degree. In particular, the horizontal and vertical sizes of the ellipsoids (dipeptides) and rhomboids (domains) were made proportional to the weighted outdegree and indegree, respectively, showcasing source-and-sink relationships. For pragmatic visualization and to avoid vanishing of 0-degree entities, the weighted degree vectors were shifted by $\div \{\text{Weighted-Degree-Mean (WDM)}\} \times 10 + 10$, where WDM = 1451 and 1688 for dipeptide and domain network, respectively. The width of arcs joining the dipeptides and domains was made proportional to the number of 'shared domains and dipeptides, respectively. Red arrowheads indicate the origin of major waves of recruitment in the time event waterfall. The arcs symbolize the flow of time (random direction for contemporary nodes).

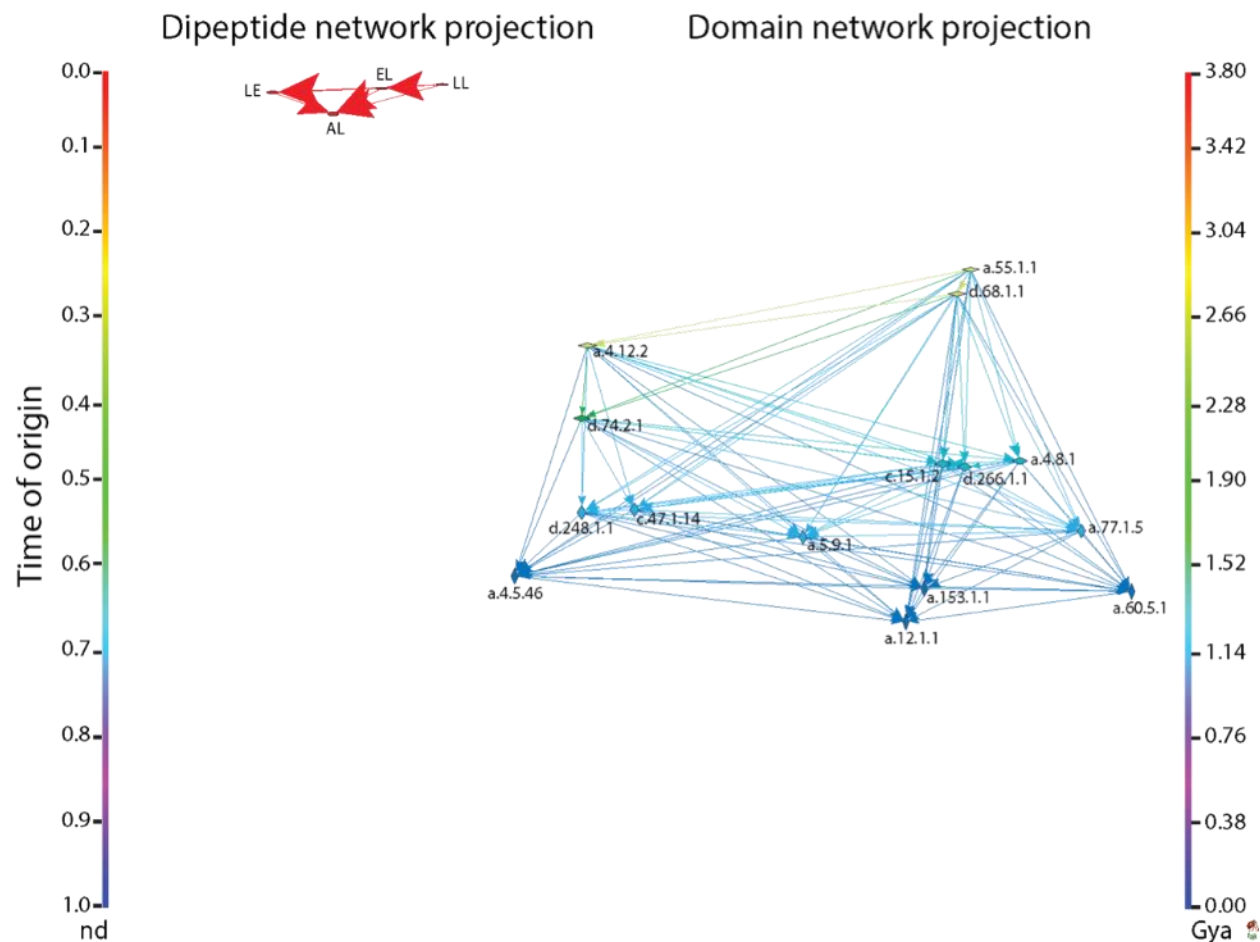
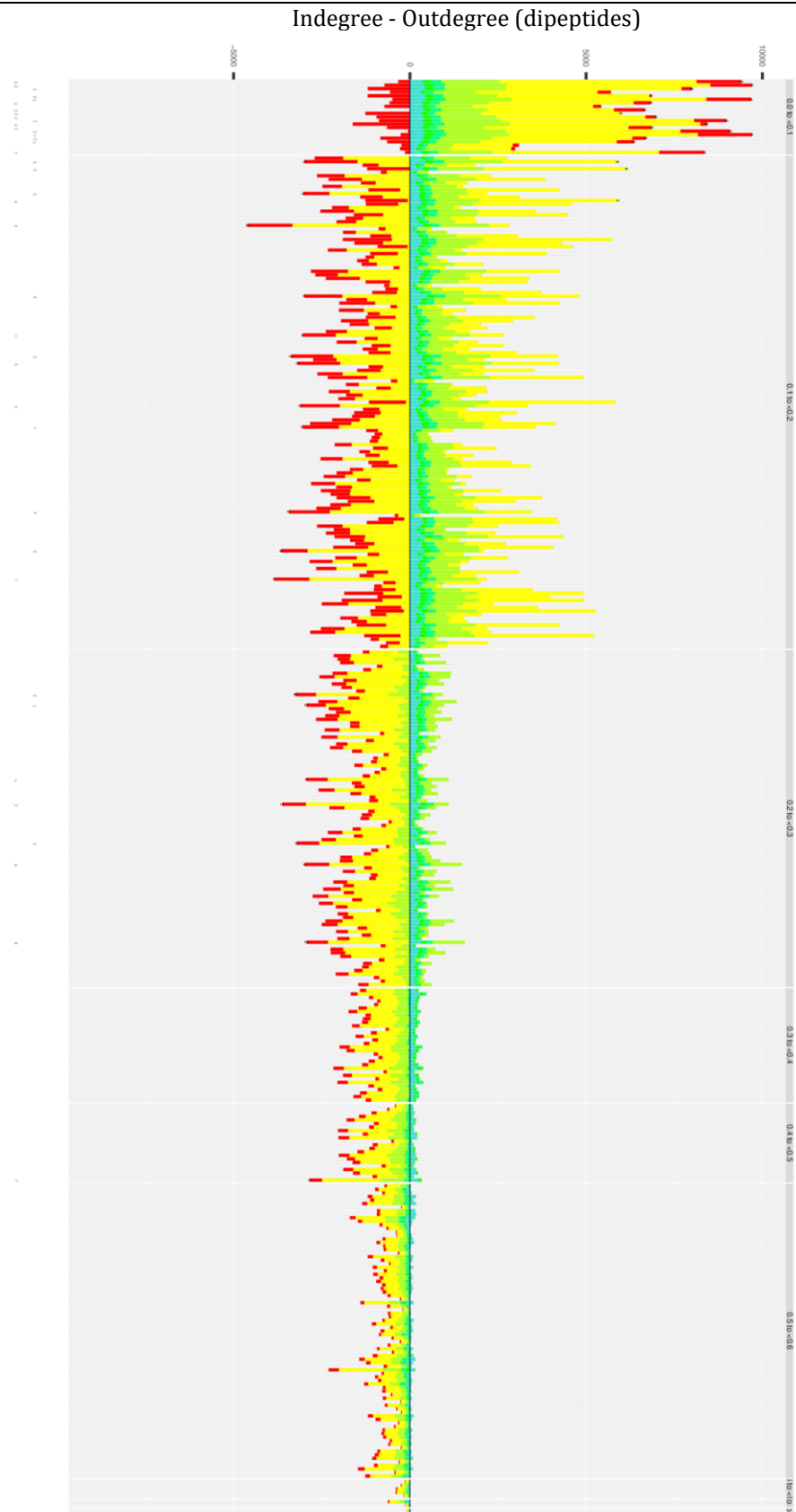


Figure 4.4. Extant D-F network projections in waterfall layout describing the evolution of dipeptides and domains with the largest (100th percentile) network connectivity. The dipeptide and domain network projections of 400 and 1,475 nodes, respectively, were each reduced with the restrictive criterion of excluding nodes with combined outdegrees and indegrees $\geq 99\%$ of those of the rest of the nodes. The set of arcs (arched arrows symbolizing flow of time) in each network was also reduced to pairing events between nodes with 100th percentile connectivity and excluded those between contemporary nodes. Nodes are arranged top-down and colored according to age (nd) on a relative 0-to-1 scale that describes evolutionary time events. Arcs are color-coded according to the age of the destination or more recent of the joined nodes. Ages are also time-calibrated with a molecular clock of domains, which uses fossils and microfossils, geochemical, biochemical, and biomarker data (Wang *et al.*, 2011). The domain origin is given as ‘billion years ago’ (Gya). The dipeptide and domain nodes were labeled with SCOP (Murzin *et al.*, 1995) descriptors. To showcase source-and-sink relationships, node symbol sizes were scaled proportional to the weighted outdegree and indegree along the horizontal and vertical axes, respectively. The original weighted degrees were shifted by $\div \{\text{Weighted-Degree-Mean (WDM)}\} \times 10 + 10$, where WDM = 1451 and 1688 for dipeptide and domain network, respectively, to include 0-degree nodes for pragmatic visualization. The modular spread of nodes was based on VOS clustering (see methods).

A)



B)

Indegree - Outdegree (domains)

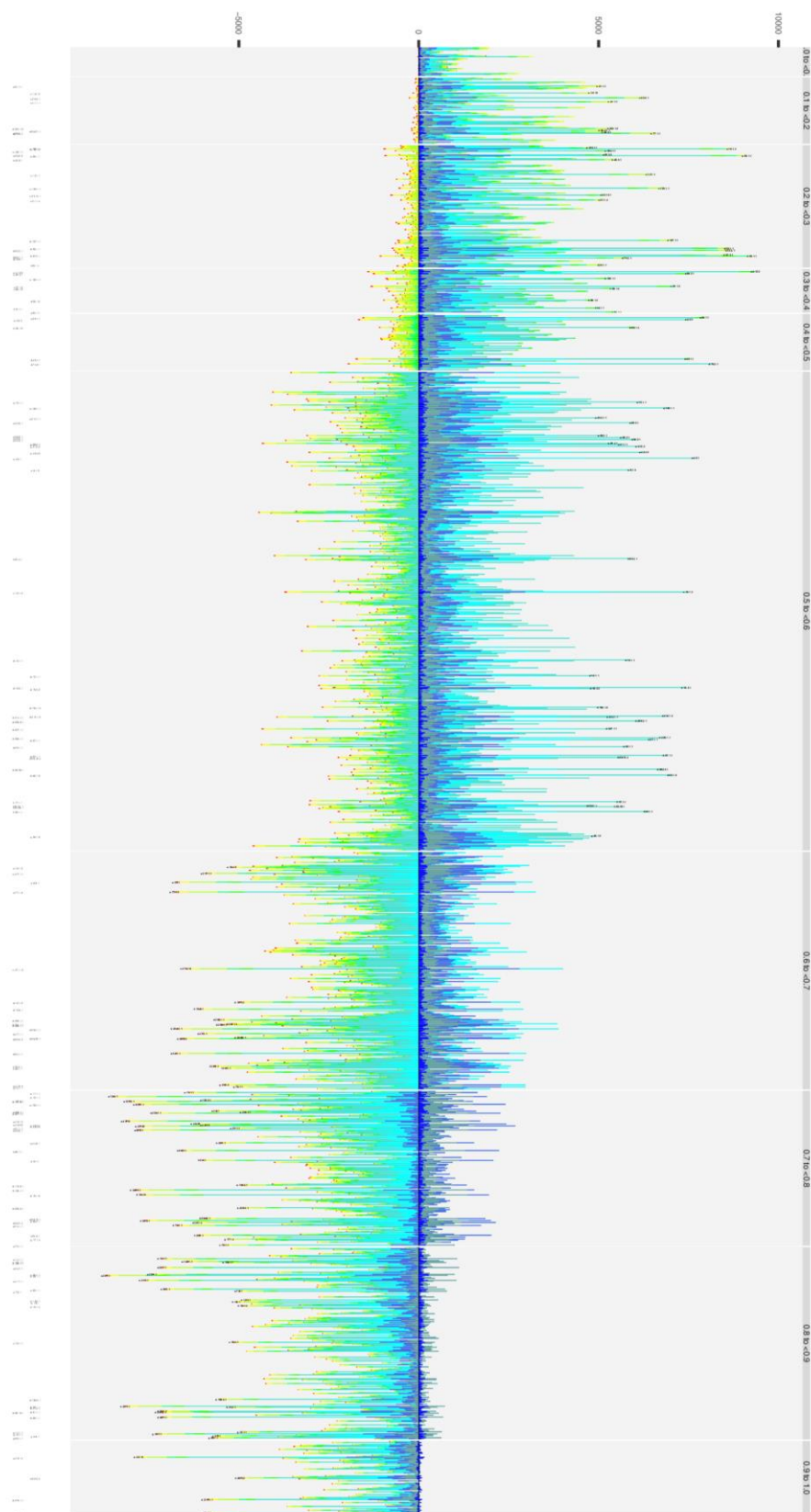


Figure 4.5. Chronological accumulation of connectivity in dipeptide and domain networks. The stacked bar charts depict the chronological accumulation of connections (arcs) in events along the timeline of dipeptide (**A**) and domain (**B**) innovation. Nodes in 95th percentile of connectivity are labeled using standard SCOP nomenclature (Murzin *et al.*, 1995). Each event corresponds to the discovery of dipeptides and domains from one of 225 events, respectively, along a timeline that spans the origin of proteins ($nd = 0$) and the present ($nd = 1$). For visualization purposes, the timeline of events was coarse-grained into 10 age bins (colored red-to-blue). For each node, the number of connections to nodes appearing earlier (indegree) or later (outdegree) in evolution were recorded and displayed as colored stacks in the stacked bars. The charts portray sink-source relationships in the recruitment of basic dipeptide footprints as viewed from the perspectives of dipeptides and domains.

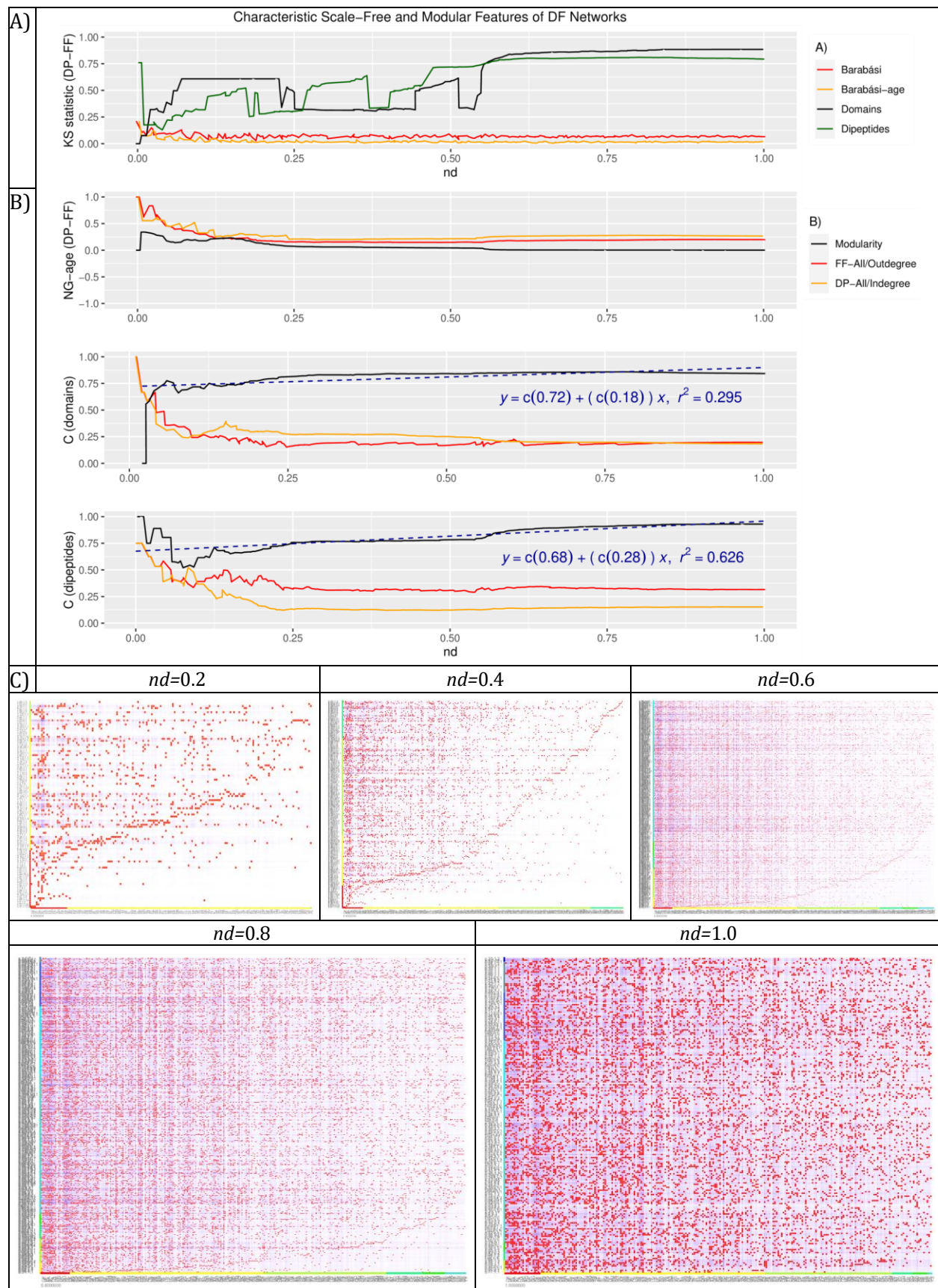


Figure 4.6. Scale-free and modular network behavior. (A) Transference of the scale-free property in the D-F (DP-FF) network. The KS fit statistic measures network-degree deviations from the fitted power law distribution. Lower KS indicates better fit (Newman, 2005; Clauset, Shalizi and Newman, 2009). The reference Barabási (red) and Barabási-Age (orange) curves are included for comparison. The generated scale-free network controls consider the preferential attachment probability of an old node to be proportional to its degree (Barabási) or to both its age and degree (Barabási-Age). (B) Modularity of growing networks. NG with default membership (partition) defined by age (NG_{age}) was computed for the D-F (DP-FF) network. NG_{age} indicates mixing of nodes by age in an assortative (≥ 0) or disassortative (< 0) manner across modules (Newman and Girvan, 2004). The average *Clustering Coefficient* (C) for dipeptide and domain networks describes the averaged ratio of the triangles to the connected triples over all nodes, where the networks are simplified (undirected/unweighted) (Wasserman and Faust, 1994; Ravasz *et al.*, 2002; Barrat *et al.*, 2004). We report the coefficients of linear regression models (navy-blue) over C for the domain network as -0.000078 by network size (N) and 0.18 by age, and those for the dipeptide network as 0.000580 by N and 0.28 by age. Linear regression lines shown are by age (nd). Normalized average degree (avg. degree) curves, computed as mean-/max-degree of the network at an event, were included as reference controls. Separate curves were computed for the ‘alldgree’ of dipeptide and domain sets of the D-F network and for the ‘outdegree’ and ‘indegree’ of dipeptide and domain networks each. Degrees were cumulative and weighted. Scores and indices were calculated for each event of the evolving networks. Age (nd) is indicated in a relative 0-to-1 scale. (C) Progression of pairwise modularity in the D-F network. The cells of the heatmaps represent modular strength between a D-F and domain as compared to their individual connectivity with the rest of the network, scaled by the absolute log10 value of the network wide modularity index NG_{age} at that event (Newman and Girvan, 2004). The first three panels illustrate the hidden switch of power law and modularity properties between dipeptides and domains. The last two panels correspond to the most-distinguishable plateaued D-F network. Network degrees were progressively reduced to 75, 50 and 25 percentiles of connectivity at time points 0.6, 0.8 and 1.0, respectively, before calculation of modularity matrices, for better visualization of significant modularity tiles. The prominent D-F units and domain structures involved in the two major waves of functional innovation along with sister hubs having red tiles are displayed using established SCOP nomenclature (Murzin *et al.*, 1995), ordered ascendingly and color-coded according to node age.

4.7 References

- Aharonovsky, E., & Trifonov, E. N. (2005). Protein sequence modules. *Journal of Biomolecular Structure and Dynamics*, 23(3), 237–242.
- Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1), 47.
- Aziz, M. F., & Caetano-Anollés, G. (2021). Evolution of networks of protein domain organization. *Scientific Reports*, 11(1), 1–18.
- Aziz, M. F., Caetano-Anollés, K., & Caetano-Anollés, G. (2016). The early history and emergence of molecular functions and modular scale-free network behavior. *Scientific Reports*, 6.
- Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509.
- Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11), 3747–3752.

- Berezovsky, I. N., & Trifonov, E. N. (2001). Van der Waals locks: loop-n-lock structure of globular proteins. *Journal of Molecular Biology*, 307(5), 1419–1426.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, 28(1), 235–242.
- Borg, I., & Groenen, P. (2003). Modern multidimensional scaling: theory and applications. *Journal of Educational Measurement*, 40(3), 277–280.
- Caetano-Anollés, G., Kim, H. S., & Mitternthal, J. E. (2007). The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proceedings of the National Academy of Sciences*, 104(22), 9358.
- Caetano-Anollés, Gustavo, Aziz, M. F., Mughal, F., & Caetano-Anollés, D. (2021). Tracing protein and proteome history with chronologies and networks: folding recapitulates evolution. *Expert Review of Proteomics*, 18(10), 863–880.
- Caetano-Anollés, Gustavo, Kim, K. M., & Caetano-Anollés, D. (2012). The phylogenomic roots of modern biochemistry: Origins of proteins, cofactors and protein biosynthesis. *Journal of Molecular Evolution*, 74(1–2), 1–34.
- Caetano-Anollés, Gustavo, Wang, M., & Caetano-Anollés, D. (2013). Structural phylogenomics retrodicts the origin of the genetic code and uncovers the evolutionary impact of protein flexibility. *PloS One*, 8(8), e72225.
- Caetano-Anollés, Gustavo, Wang, M., Caetano-Anollés, D., & Mitternthal, J. E. (2009). The origin, evolution and structure of the protein world. In *Biochemical Journal* (Vol. 417, Issue 3, pp. 621–637). <https://doi.org/10.1042/BJ20082063>
- Caetano-Anollés, Gustavo, Yafremava, L. S., Gee, H., Caetano-Anollés, D., Kim, H. S., & Mitternthal, J. E. (2009). The origin and evolution of modern metabolism. *The International Journal of Biochemistry & Cell Biology*, 41(2), 285–297.
- Caetano-Anollés, K., & Caetano-Anollés, G. (2013). Structural phylogenomics reveals gradual evolutionary replacement of abiotic chemistries by protein enzymes in purine metabolism. *PLoS One*, 8(3), e59300.
- Chung, F. R. K., Erdős, P., & Spencer, J. (1983). On the decomposition of graphs into complete bipartite subgraphs. In P. Erdős, L. Alpar, G. Halasz, & A. Saekőzy (Eds.), *Studies in pure mathematics, To the memory of Paul Tur{á}n* (pp. 95–101). Budapest: Bikh{VH a}user Verlag.
- Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6), 66111.
- Clauset, Aaron, Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4), 661–703.
- Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5), 1–9.

- Debès, C., Wang, M., Caetano-Anollés, G., & Gräter, F. (2013). Evolutionary optimization of protein folding. *PLoS Comput Biol*, 9(1), e1002861.
- Delaney, W., & Vaccari, E. (1989). *Dynamic models and discrete event simulation*.
- Diestel, R. (2010). *Graph theory. Graduate Texts in Mathematics* (Fourth Edi, Vol. 173). Heidelberg: Springer-Verlag.
- Goncearenco, A., & Berezovsky, I. N. (2010). Prototypes of elementary functional loops unravel evolutionary connections between protein functions. *Bioinformatics*, 26(18), i497--i503.
- Gough, J., Karplus, K., Hughey, R., & Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of Molecular Biology*, 313(4), 903–919.
- Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3), 299–314.
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., & Barabási, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804), 651–654.
- Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12), 2577–2637.
- Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1), 7–15.
- Kim, H. S., Mittenthal, J. E., & Caetano-Anollés, G. (2013). Widespread recruitment of ancient domain structures in modern enzymes during metabolic evolution. *Journal of Integrative Bioinformatics*, 10(1), 214.
- MacDougall, M. H. (1987). *Simulating computer systems: Techniques and tools*.
- Mayer, C. (2022). Spontaneous Formation of Functional Structures in Messy Environments. *Life*, 12(5), 720.
- Mittenthal, J. E., Caetano-Anollés, D., & Caetano-Anollés, G. (2012). Biphasic patterns of diversification and the emergence of modules. *Frontiers in Genetics*, 3, 147.
- Mrvar, A., & Batagelj, V. (2016). Analysis and visualization of large networks with program package Pajek. *Complex Adaptive Systems Modeling*, 4(1), 1–8.
<http://casmodeling.springeropen.com/articles/10.1186/s40294-016-0017-8>, Date of access: 08/02/2016
- Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification*, 31(3), 274–295.

- Murzin, A. G., Brenner, S. E., Hubbard, T., Chothia, C., & others. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4), 536–540.
- Newman, M E J, & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2), 26113.
- Newman, Mark E J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5), 323–351.
- Newman, Mark E J, Strogatz, S. H., & Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2), 26118.
- Overbeek, R., Larsen, N., Pusch, G. D., D'Souza, M., Selkov Jr, E., Kyrpides, N., Fonstein, M., Maltsev, N., & Selkov, E. (2000). WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Research*, 28(1), 123–125.
- Pidd, M. (2006). Computer simulation in management science. *JOURNAL-OPERATIONAL RESEARCH SOCIETY*, 57(3), 327.
- R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 2013. ISBN 3-900051-07-0.
- Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabási, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586), 1551–1555.
- Strogatz, S. H. (2001). Exploring complex networks. *Nature*, 410(6825), 268–276.
- Tanaka, R. (2005). Scale-rich metabolic networks. *Physical Review Letters*, 94(16), 168101.
- Van Eck, N. J., & Waltman, L. (2007). VOS: a new method for visualizing similarities between objects. In *Advances in Data Analysis: Proceedings of the 30th Annual Conference of the German Classification Society* (pp. 299–306). Heidelberg: Springer Verlag.
- Wagner, A., & Fell, D. A. (2001). The small world inside large metabolic networks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1478), 1803–1810.
- Waltman, L., van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, 4(4), 629–635.
- Wang, G., & Dunbrack Jr, R. L. (2003). PISCES: a protein sequence culling server. *Bioinformatics*, 19(12), 1589–1591.
- Wang, M, & Caetano-Anollés, G. (2009). The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. *Structure*, 17(1), 66–78.
- Wang, Minglei, Jiang, Y.-Y., Kim, K. M., Qu, G., Ji, H.-F., Mittenthal, J. E., Zhang, H.-Y., & Caetano-Anollés, G. (2011). A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation. *Molecular Biology and Evolution*, 28(1), 567–582.

Wang, Minglei, Kurland, C. G., & Caetano-Anollés, G. (2011). Reductive evolution of proteomes and protein structures. *Proceedings of the National Academy of Sciences*, 108(29), 11954–11958.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. 8.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684), 440–442.

CHAPTER 5: CONCLUSIONS

This study investigates patterns of emergence of hierarchical modularity of structural and functional cooption and pairwise recruitment through exploration of network connectivity in the domain-multidomain, loop-domain, and dipeptide-dipeptide mappings. We trace the coevolutionary history of the domains and domain combinations of proteins, defined at the fold superfamily (FSF) level of SCOP classification (Murzin *et al.*, 1995), loop prototypes (loops) as defined in ArchDB (Bonet *et al.*, 2014), and dipeptides. We unfold the evolutionary mechanisms and the rationale driving the history of networks from the perspectives of protein domain organization, loops, dipeptides. It also analyzes the dynamics leading to an adaptation of the structure of these networks to scale-free properties. We also explain the underlying processes responsible for such network structure. To conduct this investigation, we highlight the cooption of protein domains as they interact with other domains throughout a timeline of protein evolution. For this purpose, a subset of 6,162 protein domains and multidomains was taken into consideration. Domains are the structural, functional and evolutionary units of proteins. They combine to form multidomain proteins. The evolutionary history of this molecular combinatorics has been studied with phylogenomic methods. Here, we construct networks of domain organization and explore their evolution. Our network model consists of five distinct composition-based and topology-based ‘operational’ criteria for defining nodes and links of the networks: 1) domain and multidomain (domain organization) composition, 2) domain and multidomain pairs decomposition, 3) domain and multidomain adjacent-pairs decomposition, 4) splicing of domain pairs from domain organizations, and 5) splicing of adjacent domain pairs from domain organizations, which we carefully define. The evolutionary mappings (based on the ages) of the protein domain

organization and growing interactions revealed the history of an evolving biological network. This evolutionary timeline stemmed from a phylogenomic analysis (Wang and Caetano-Anollés, 2009) and was demarcated by 169 unique time events. In case of loop prototype networks, 1,937 loops and 1,442 fold families (domains) were mapped with 228 unique time events. Also, abundance data of course 400 dipeptides for domains and genomes was considered over 225 unique time events. These time events (or time slivers) were mapped onto the evolving networks to depict evolving network configurations from the origin of proteins to the modern protein world. These time-varying (evolving) networks revealed two ancient waves of structural novelty arising from ancient ‘*p*-loop’ and ‘winged helix’ domains and a massive ‘big bang’ of domain organization. The evolutionary recruitment of domains was highly modular, hierarchical, and ongoing. Domain rearrangements elicited non-random and scale-free network structure. Comparative analyses of preferential attachment, randomness and modularity of networks showed yin-and-yang complementary transition patterns along the evolutionary timeline. Remarkably, evolving networks highlighted a central evolutionary role of cofactor-supporting structures of non-ribosomal peptide synthesis (NRPS) pathways, likely crucial to the early development of the genetic code. Some highly modular domains featured dual response regulation in two-component signal transduction systems with DNA-binding activity linked to transcriptional regulation of responses to environmental change. Interestingly, hub domains across the evolving networks shared the historical role of DNA binding and editing, an ancient protein function in molecular evolution. Our investigation unfolds historical source-sink patterns of evolutionary recruitment that further our understanding of protein architectures and functions. We here present the first tracing of evolutionary ages of protein domain architectures in evolving networks of domain organization. These growing networks revealed major evolutionary pathways of molecular

recruitment of domains and functions. Two prominent ancestral waves of structural novelty involved ancient domain innovations and founder ‘*p*-loop’ and ‘winged helix’ domain structures. We found that the evolutionary recruitment of domains and multidomains in proteins is ongoing and highly modular. Remarkably, the networks highlighted the role of cofactor-supporting structures of NRPS pathways, which were backbone to the early evolution of the genetic code. The evolving domain rearrangements featured multitier evolutionary episodes of scale-free network structure, hierarchy, and modular behavior. Remarkably, our analyses support biphasic patterns of diversification and module emergence that we have observed earlier (Mittenthal, Caetano-Anollés and Caetano-Anollés, 2012; Aziz, Caetano-Anollés and Caetano-Anollés, 2016). In an initial phase, at the cusp of architectural diversification, the modular components of emerging domain organization associated through weak linkages of recruitment. The second phase was massive and prolonged, with a multitude of modules appearing after the ‘big bang’ of the protein world, supporting the onset of organismal diversification. Thus, the origin and evolution of structural domains and modular and non-modular loops through ‘waves’ of recruitment phases is backed by strong genomic and structural evidence, which as we showcased is feasible to be used in modelling of the gradual evolution of proteins, illustrated here with AlphFold2-modelled P-loop hydrolases.

We also traced evolutionary age onto growing elementary functionome networks and their projections verified the two waves of functional innovation in elementary functional loops involving founder ‘*p*-loop’ and ‘winged helix’ domain structures (Aziz, Caetano-Anollés and Caetano-Anollés, 2016). We found that the dynamics of recruitment of loop motifs into domain structures is modern and highly modular. The emergence of molecular functions featured hierarchical modularity and a continuing power law behavior. The EF network showcases

integrated communities of interacting structural parts defining modular classes of molecular functions. Our analyses suggest that this hierarchical modularity evolved through a biphasic process of diversification³⁸. The biological functions of proteins are defined by conserved amino acid sequence motifs embedded into the structural scaffolds of protein domains. Thus, we use a bipartite network approach to study the history of how these elementary functional loop prototype (loop) prototypes combine to form domain structures defined at the fold family (domain) level of SCOP classification. The resulting ‘elementary functionome’ (EF) network and its loop motif and domain domain graph projections described evolutionary ‘waterfalls’ emphasizing the emergence of primordial functions. Timed waterfalls revealed historical sharing of ancient functional loop motifs by domain structures in two primordial phases of functional innovation that involved founder ‘*p*-loop’ and ‘winged helix’ domains. In phase 1, the components of the emerging functionome (loop motifs and domain structures) associated enormously, but, with weak linkages, through processes of recruitment. These motifs and structures then diversified, engaged in competition, and were selected for functional performance. Useful emerging interactions constrained their associations. This caused tight linkages to self-organize into modules and express as closely-knit communities. In a relatively longer phase 2, variants of the modules evolved and instigated a new generative cycle of higher-level organization highlighted with scale-free module recruitment that is modern and ongoing. We also uncover a dual hourglass recruitment pattern of embedding of modular motifs in structures, exchanging ‘preferential’ cooption properties between ancient loops and emerging domains. Notably, we found the emergence of molecular functions to be modernistic and ongoing. Networks of loop motifs and domain structures expanded metabolic pathways and translation functionalities while simultaneously fostering hierarchical modularity in network evolution. We have previously shown that biphasic patterns are prevalent in the size,

dipeptide makeup, and loop-mediated flexibility of proteins^{1,8}, which is likely linked to their intrinsic disorder. This adaptation to a biphasic pattern of change seems to be a general phenomenon in network biology.

Finally, we examined the interactome of 400 dipeptides with 1,475 SCOP fold families and 1,561 proteomes for modular and scale-free properties. This bipartite network that resulted in dipeptide, domain and genome network projections helped examine how dipeptide behave as building blocks of domains and in establishing an evolutionary ‘dipeptide footprint’. Ages of dipeptides were directly borrowed from the associated domains. We tested the hypotheses of whether collective dipeptide frequencies can be significantly modeled as the proteome size and can the collective normalized dipeptide frequencies be significantly modeled as the dipeptide age. We also examined how dipeptide abundances relate to proteome size and dipeptide age in factorial design. At last, we investigate if the domain and genome-sharing based dipeptide networks indicate preferential attachment. Most properties were evidently significant and rejected the null hypotheses, with networks turning out to be scale-rich instead of scale-free. We anticipate similar bowtie-like hierarchical modularity patterns emerging from the recruitment of dipeptides as building-blocks of domains.

Hence, these biphasic patterns observed in both functional and structural molecular histories are prevalent in biology and impact size, dipeptide makeup, and loop-mediated flexibility of proteins, possibly due to their intrinsic disorder(Caetano-Anollés, Wang and Caetano-Anollés, 2013; Debès *et al.*, 2013). These patterns seem to favor the ancient and pre-modern entities to form new combinations (by fusion) or to fragment into new structures (by fission) later in evolution. In turn,

the domains, and loops of the middle period of the timeline struggle for dominance. This is specifically evident when we dissect two prominent aspects of the evolutionary process of recruitment, the ability of domains, loops, and dipeptides to act as 'donors' and 'acceptors'. These networks were also evaluated using scale-free, random, and modular network models to interpret the gradual evolutionary selection of one model in favor of the other. We propose that biphasic patterns in evolving networks are integral to biological history. The history of these networks was visualized using Pajek (Mrvar and Batagelj, 2016) to reveal distinct patterns of connectivity among domains and multidomains, loops and domains, and dipeptides, and highlight variations of inherent characteristics of domains along the timeline. Particularly, we discover an interesting pattern of highly concentrated connectivity associated to the domains and loops of either the earliest or the post-middle era of domain evolution, while sparse connectivity exhibited by the domains of the middle period of the evolutionary timeline. This observation provides evidence for a biphasic (Mittenthal, Caetano-Anollés and Caetano-Anollés, 2012) or hourglass (Caetano-Anollés *et al.*, 2011) pattern of connectivity in protein domain, loop and dipeptide networks.

APPENDIX A: LIST OF ACRONYMS

ANOVA	Analysis of Variance
ASCII	American Standard Code for Information Interchange
C	average Clustering Coefficient
C-ratio	Clustering ratio
CWDN	Cumulative Weighted Degree per Node
DD	Degree Dispersion
DES	Discrete Event Simulation
domain	Fold Family
FGC	Fast Greedy Community detection algorithm
FSF	Fold SuperFamily
GLM	Generalized Linear Model
Gy	Geological years (billions of years)
Gya	Geological years ago (billions of years ago)
KS	Kolmogorov-Smirnov statistical test of power law fit
LC	Loop Classifier
EF	Elementary Functionome
LM	Linear Model
Log L	log Likelihood
MANOVA	Multivariate Analysis of Variance
N	Network size
ND	Node Distance

NEA	Network Age
NG	Newman-Girvan modularity algorithm
NG _{age}	NG partitioned by age
NG _{vos}	NG partitioned by VOS clustering
NOA	NODE Age
NTP	Nucleoside Triphosphate
PC	Principal Component
PCA	Principal Components Analysis
PDB	Protein Data Bank
SCOP	Structural Classification of Proteins
VOS	Visualization of Similarity clustering algorithm
VQ	VOS Quality index

APPENDIX B: SUPPLEMENTARY INFORMATION FOR CHAPTER 2

B.1 Supplementary Text

B.2 Supplementary Materials and Methods

B.3 Supplementary Figures

B.4 Supplementary Tables

B.5 Supplementary Videos

Supplementary Video 1. Simulation of evolutionary growth of the SPX network of domain organization from ancient to modern times of the protein world, March 18, 2021.

Supplementary Video 2. Simulation of *waterfall* model of the SPX network based on the linear (incremented by 1) progression of networks with nodes representing connectivity higher than or equal to various (0 to 100) percentiles of combined outdegree and indegree at Age 1, March 18, 2021.

Supplementary Video 3. Simulation of pairwise NG_{age} modularity of the SPX network through progression of heat maps and dendrograms over the timeline, March 18, 2021.

B.6 Supplementary Data Files

Supplementary Data File 1. Data matrix used to build the phylogenomic tree of architectures. Genome identifiers are described in Table S1 of ref.(Wang, Kurland and Caetano-Anollés, 2011), March 18, 2021.

Supplementary Data File 2. Most parsimonious phylogenomic tree of domain architectures, March 18, 2021.

Supplementary Data File 3. Most parsimonious phylogenomic tree of domain architectures, with taxa labeled with SCOP ccs identifiers, March 18, 2021.

More supplementary information accompanies this chapter at <http://www.nature.com/srep> and <https://github.com/gcalab/SciRep>.

APPENDIX C: SUPPLEMENTARY INFORMATION FOR CHAPTER 3

C.1 Supplementary Materials and Methods

C.2 Supplementary Figures

C.3 Supplementary Videos

Supplementary Video 1. Simulation of pairwise NG_{age} modularity of the EF network through progression of heat maps over the timeline, Nov 26, 2021.

Supplementary Video 2. Simulation of hierarchical organization of the EF network based on NG_{age} modularity through progression of dendrograms over the timeline, Nov 26, 2021.

Supplementary Video 3. Simulation of *waterfall* model of the loop network projection based on the linear (incremented by 1) progression of networks with nodes representing connectivity higher than or equal to various (0 to 100) percentiles of combined outdegree and indegree at Age 1, March 17, 2022.

Supplementary Video 4. Simulation of *waterfall* model of the domain network projection based on the linear (incremented by 1) progression of networks with nodes representing connectivity higher than or equal to various (0 to 100) percentiles of combined outdegree and indegree at Age 1, March 17, 2022.

Supplementary Video 5. Simulation of network growth with progression of node sizes and connectivity in the EF bipartite network over the evolutionary timeline, June 02, 2022.

Supplementary Video 6. Simulation of network growth with progression of node sizes and connectivity in the loop projection network over the evolutionary timeline, June 02, 2022.

Supplementary Video 7. Simulation of network growth with progression of node sizes and connectivity in the domain projection network over the evolutionary timeline, June 02, 2022.

APPENDIX D: SUPPLEMENTARY INFORMATION FOR CHAPTER 4

D.1 Supplementary Materials and Methods

D.2 Supplementary Figures

D.3 Supplementary Videos

Supplementary Video 1. Simulation of *waterfall* model of the dipeptide network projection based on the linear (incremented by 1) progression of networks with nodes representing connectivity higher than or equal to various (0 to 100) percentiles of combined outdegree and indegree at Age 1, July 13, 2022.

Supplementary Video 2. Simulation of *waterfall* model of the domain network projection based on the linear (incremented by 1) progression of networks with nodes representing connectivity higher than or equal to various (0 to 100) percentiles of combined outdegree and indegree at Age 1, July 13, 2022.

Supplementary Video 3. Simulation of pairwise NG_{age} modularity of the D-F network through progression of heat maps over the timeline, July 13, 2022.

Supplementary Video 4. Simulation of hierarchical organization of the D-F network based on NG_{age} modularity through progression of dendrograms over the timeline, July 13, 2022.