THE EARLY HISTORY AND EMERGENCE OF MOLECULAR FUNCTIONS AND MODULAR SCALE-FREE NETWORK BEHAVIOR

BY

MUHAMMAD FAYEZ AZIZ

THESIS

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

Urbana, Illinois

Master's Committee:

Professor Gustavo Caetano-Anollés, Chair
Professor Sandra Luisa Rodriguez-Zas
Associate Professor Kaustubh Bhalerao
Assistant Professor Jian Ma

ABSTRACT[1]

The formation of protein structural domains requires that biochemical functions, defined by conserved amino acid sequence motifs, be embedded into a structural scaffold. Here we trace domain history onto a bipartite network of elementary functional loop (EFL) sequences and domain structures defined at the fold superfamily (FSF) level of Structural Classification of Proteins (SCOP). The resulting 'elementary functionome' network and its EFL and FSF graph projections unfold evolutionary 'waterfalls' describing emergence of primordial functions. Waterfalls reveal how ancient EFLs are shared by FSF structures in two initial waves of functional innovation that involve founder '*p*-loop' and 'winged helix' domain structures. They also uncover a dynamics of modular motif embedding in domain structures that is ongoing, which transfers 'preferential' cooption properties of ancient EFLs to emerging FSFs. Remarkably, we find that the emergence of molecular functions induces hierarchical modularity and power law behavior in network evolution as the networks of motifs and structures expand metabolic pathways and translation.

[1] Materials in this document (all chapters, including the abstract) are reprinted, with permission, from Aziz MF, Caetano-Anollés K and Caetano-Anollés G, 2015, "The early history and emergence of molecular functions and modular scale-free network behavior," *manuscript submitted*

To Asma, Rania and Rayan

ACKNOWLEDGMENTS

This project would not have been possible without the support of my adviser, Prof. Dr. Gustavo Caetano-Anollés. I thank him for multiple revisions and feedback, inspiration and confidence, and making sense out of confusion. I also thank my committee members, Prof. Jian Ma and Prof. Sandra Rodriguez, for providing helpful suggestions, appreciation and support. Thanks also need to be extended to Dr. Minglei Wang, for the base data set used to generate the timelines, and to Kelsey Caetano-Anollés, for pioneering the generation of primary data set and initial network analysis. I would also like to acknowledge the funding support for this research provided by the National Science Foundation and the United States Department of Agriculture, and initial support from the COMSATS Institute of Information Technology, Pakistan. Thanks to my father for imparting his knowledge and transferring his skills to me, my mother for the nourishment, building my intellectual base and affection, and my brothers for the inspiration and guidance with their own doctoral degrees. I thank my wife for sticking with me in thick and thin, and of course for her love and beauty, and my children for cheering me up in my down moments. Finally, all thanks to God (Allah) Almighty for everything.

TABLE OF CONTENTS

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*

In order to explain the structural and functional complexities of the protein world, protein domain structure must emerge from prior structural states and must fulfill the evolutionary principle of spatiotemporal continuity. We recently argued that these prior states involve the combination of dipeptides to form three-dimensional loop structures and that these non-regular structures provide the necessary flexibility to unfold molecular functions and genetics (Caetano-Anollés et al. 2013). Using a phylogenomic framework, we previously reconstructed evolutionary timelines of molecular accretion in which molecules acquire substructural or modular parts in their molecular makeup. These timelines are directly generated from the sequence and structure of thousands of nucleic acid molecules and millions of protein sequences encoded in hundreds of genomes. For proteins, timelines make explicit the gradual evolutionary appearance of protein domain structures and molecular functions (reviewed in Caetano-Anollés et al. 2009a, 2012; Kim and Caetano-Anollés 2010; Edwards et al. 2013) and their combinatorial rearrangement in proteins (Wang and Caetano-Anollés 2009). They also allow the evolutionary tracing of chemical and biophysical properties. For example, tracing chemical mechanisms in enzymatic reactions uncovered the natural history of biocatalysis (Nath et al. 2014). Similarly, tracing contact order (i.e. localization of amino acid contacts in protein's tertiary structure),

which is correlated to flexibility, showed that folding speed is optimized and increases in protein evolution (Debès et al. 2013). In dynamic metabolomics networks of *Escherichia coli*, subjection to stress stochastically induces biphasic-rewiring and modularity at regular time intervals of few minutes (Aziz et al. 2012).

The formation of domains from dipeptide constituents must also involve stable intermediates that would act as scaffolds of the flexible functional loops of emerging structures smaller than the size of an average compact domain (~100 amino acid residues in length) (Trifonov and Frenkel 2009; Sobolevsky et al. 2013). These intermediate prior forms have been postulated to be small peptides (~25-30 residues) forming closed loops stabilized by van der Waals locks (Berezovsky and Trifonov 2001; Aharonovsky and Trifonov 2005). Their history has been traced back to few prototypes that are universally present in structures and are believed to be modern determinants of molecular function (Berezovsky et al. 2000; Goncearenco and Berezovsky 2010). In a recent study, distant evolutionary connections of these 'elementary functional loops' (EFLs) revealed patterns of motif reuse in archaeal proteins (Goncearenco and Berezovsky 2012). Here we map the coevolutionary history of EFL prototypes and protein domains defined at fold superfamily (FSF) level of structural abstraction of the Structural Classification of Proteins (SCOP) database (Murzin et al. 1995). In SCOP, proteins with similar 3D arrangements of secondary structures exhibiting identical topological connections have been classified as folds. Within folds, proteins whose structure and functional features indicate a common evolutionary origin are categorized as FSF groups. FSFs with more than 30% amino acid residues similarity are further classified into fold families (Figure 1.1). Mining EFL interactions at FSF level reveals remarkable patterns of emergence of molecular functions that are likely of very ancient origin.

CHAPTER 2: MATERIALS AND METHODS

2.1 FSF and EFL prototype data

FSF domain structures were defined according to SCOP version 1.75 (Murzin et al. 1995). The relative ages, calculated as node distance (*nd*) values from FSF phylogenetic trees, of first appearances of FSFs were obtained from a previously published timeline of protein domain evolution (Wang et al. 2011). The timeline was derived directly from a phylogeny describing the evolution of 1,730 FSFs reconstructed from a census of domain structure in 749 genomes of 52 archaeal, 478 bacterial and 219 eukaryal organisms (dataset A749). A calibrated molecular clock of FSF structures ($t = -3.831nd + 3.628$) was used to calculate geological age in billions of years (Gy) (Wang et al. 2011). EFL prototypes were previously identified computationally in the complete genome sequences of 68 archaeal organisms by iteratively deriving sequence profiles with a scoring function that weighs profile positions according to information content followed by hierarchical clustering (Goncearenco and Berezovsky 2010). Since EFLs are embedded in FSF structure and both EFLs and FSFs describe functional and structural abstractions, the age of FSFs can be directly transferred to EFLs. We used two likely schemes to do this: (i) the age of the EFL is the age of the most ancient associated FSFs, or (ii) the age of the EFL is the age of the most recent of the most ancient couple of associated FSFs (or the age of the single associated FSF). The mappings consider the age of an EFL as either the age of the first structural scaffold or the age when the EFL function is first transferred between structural scaffolds, respectively. Since both schemes provide similar mappings, we only show mappings derived using the second more conservative scheme. The strongest 43 EFLs out of 138 profiles were selected and mapped

against non-redundant FSF domains of the SCOP 1.75 database in order to establish evolutionary connections with FSFs responsible for molecular functions (Goncearenco and Berezovsky 2010).

2.2 Network visualization and analysis

Networks were visualized and analyzed using Pajek (Batagelj and Mrvar 1998) 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 et al. 2010). Network properties were analyzed with graphing code constructs and packages 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 graphs, compute power law statistics and modularity indices, and construct waterfall diagrams follows.

2.2.1 Network data analyses. The EF network and the EFL and FSF network projections were visualized and analyzed using Pajek (Batagelj and Mrvar 1998) as they unfolded in the evolutionary timeline. The nodes (vertices) and links (arcs/edges) of the EF bipartite graph and its projections were encoded in Excel (ASCII) project files for network visualization. The quantitative network property of cumulative weighted degree per node (*cwdn*) was generated using custom Pajek macros. *cwdn* was compiled as three types of data matrices for each network (BOX 1). The data rows of matrices defined EFLs and FSFs, and were sorted in ascending order of node-(individual *nd*) or network-(event *nd*) age. Separate matrices were organized for the 'in' and 'out' degree types as well as each portion of the bipartite node set.

BOX 2.1. Types of data matrices

Matrix type	Matrix objective
By 'node age' (NOA)	Used for box plots and x-y line plots. Categorical columns were ordinal number, age bin, age and node label of EFL and FSF nodes in the networks. Additional columns described *cwdn* in increasing order of events. Rows were sorted by node.
By 'network age' (NEA)	Used for power law distribution graphs. They are essentially transpositions of NOA data. Columns were ordinal number, age bin, and age of networks, followed by *cwdn* arranged by nodes. Rows were sorted by events.
By 'degree dispersion' (DD)	Used for stacked bar charts. Column and row order are the same as NOA data. However, columns provided the distributions of final *cwdn* (i.e. at *nd*=1) across connected node age bins.

2.2.2 Network visualization. A set of Pajek menu commands was used to generate visualization attributes and layouts (BOX 2). Node categories were made distinguishable using various shape and color options. Evolutionary patterns were unfolded by color-coding the age of EFL and FSF nodes. A color palette was used that ranged from red for the most ancient node (*nd* = 0) to orchid for the most recent node (*nd* = 1).

BOX 2.2. Network visualization tools and commands

Pajek tool and command	Output
Network: Create Vector: Centrality: Weighted Degree: 'All', 'Output\Input'	The weighted degree vectors for undirected and directed networks, respectively.
Draw: 'Network + Vector'	Visualization of the weighted degree of nodes as node size.
Network: Create Partition: Communities: VOS clustering: 'Multi-level coarsening + Multi-level refinement'	The community-based layout of the networks, using the VOS clustering method (Van Eck and Waltman 2007; Waltman et al. 2010). In addition, modularity indices were obtained. Default parameters were used.
Draw: 'Network + Partition' and Draw: Layers: 'In y Direction'	Vertical arrangement of nodes according to their age in both bipartite and waterfall layouts of the networks with age mappings.
Draw: Layout: Energy: Kamada-Kawai: 'Optimize inside clusters only'	Visual distribution of partitions (clusters or communities) in waterfall layout, manually separated to refine the most energetically favored network configurations.

2.2.3 Charts and graphs. Graphing code constructs and packages of R (Ihaka and Gentleman 1996; R Core Team 2014) were used to visualize the *cwdn* (BOX 3).

BOX 2.3. Graphing and charting constructs and operations

R *package*: function	Result
Reshape: *	Transform textual data tables to vector form.
Lattice: *	Generate panels of network- and node-age bins.
LatticeExtra, grid: *	Resize panels.
LatticeExtra: 'panel.lmline ()'	Draw linear regression model lines and equations, e.g. in *log-log* graphs.
Data.table: *	Edit data tables, e.g., to remove empty age bins.
igraph: *	Read network graph files to calculate modularity and apply the power law distribution model to calculate its fit statistics.
R base: 'bwplot ()'	Box and whisker's plots based on NOA files, to depict measures of central tendency of *cwdn* over network age.
Lattice: 'xyplot ()'	XY scatter and line plots based on NOA files drawing final *cwdn* attained at network age 1.
R color palette: *	The color-coding of data points by node age.
ggplot2: 'ggplot ()', 'geom bar ()' and 'grid.draw ()'	Produce degree dispersion stacked bar charts and graphical trend curves of power law behavior and modularity, based on DD, NEA data and VOS modularity report files, respectively.
ggplot2: 'scale_fill_manual ()'	Color-coded stacked bars representing distribution of final *cwdn* over ages of connected node set.
ggplot2: 'layer ()'	Stack negative valued bars in reflection.
ggplot2: 'theme ()'	Color-coding of multiple curves of power law statistics and modularity indices.
gridExtra: 'grid.arrange ()' and 'arrangeGrob ()'	Plotting together of curves for comparison.
ggplot2: 'geom_smooth ()'	Addition of linear regression lines to some modularity plots.
ggplot2: 'geom_tile ()', 'scale_fill_gradient ()', 'scale_x/y_discrete ()' etc.	Customized modularity heat maps.
ggdendro: *	Plotting of dendrograms.

2.3 Power law network behavior

Scale free network behavior was studied using *P(k)* vs. *k* (i.e. probability of having *k*-neighbors vs. *k*) and *log-log* (i.e. log of *P(k)* vs. log of *k*) mappings, with linear regression models to derive γ of the power law and the determination coefficient (R^2). γ is the negative of the 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) the exponent of the fitted power law distribution (α), which assumes $P(X=x)$ is proportional to $x^{-\alpha}$; (ii) KS fit statistic, which compares the fitted distribution with the input degree vector; and (iii) the KS *p*-value, with the null hypothesis of data being drawn from the power law distribution (Newman 2005; Clauset et al. 2009). Higher α, smaller KS fit scores, and larger KS *p*-values ($\geq$ 0.05) suggest better fit to power law distributions. We also determined the maximum log likelihood of the fitted parameters and if the power law fit pattern was continuous. Reference networks were created using 'barabasi' methods of *igraph* to simulate power law and extended age-dependent graph models for the corresponding networks.

2.3.1 Barabási reference networks. Reference networks for comparative analysis of power law and modularity were generated using 'barabasi' methods of the R's *igraph* package (Barabási and Albert 1999). The ideal power law model was implemented with '`barabasi.game ()`'. An extended model was implemented with '`aging.prefatt.game ()`'. This model simulated the scale-free evolution of random graphs by altering the probability of preference of an old vertex growing multifold, exponentially with age. Networks of three sizes, 120, 82 and 38, were created to simulate reference controls for the corresponding EF, FSF and EFL networks, with properties of directionality directly transferred. Ages were assigned to individual nodes in incremental order to keep age proportion per event consistent with the timelines of the test networks.

2.3.2 Power law statistics. *P(k)* vs. *k* and *log-log* plots provided preliminary insight into the scale free behavior of a network. Power law behavior usually manifests in degree distributions exhibiting long tails (similar to *Poisson*). R libraries and operations were utilized to run the analysis (BOX 4). The distributions were color-coded. *Log P*(*k*) vs. *log k* scatter plots supplemented analysis of power law behavior.

BOX 2.4. R operations used in power law analysis

R *package*: function	Result
Lattice: '`xyplot ()`'	Plot distributions generated from NEA files.
Data.table: '`table ()`'	Compute *k* frequency tables for *P(k)* vs. *k* plots.
R base: '`length ()`'	Calculate length of degree table in order to determine *P(k)* through quotient of *k* frequencies by maximum degree.
R base : '`ppois()`'	Compute *Poisson* (*k*) to be plotted against *k*.
R base: '`log ()`'	Implement *log P(k)* vs. *log k* scatter plots.
Lattice: '`lm ()`' and '`panel.abline ()`'	Generate and draw linear regression models.
R base: '`summary (lm)$ coefficients`'	Retrieve the γ-slope power law coefficient and R^2 determination coefficient. *lm* being the linear model.
igraph: '`power.law.fit ()`'	Obtain additional statistics supporting the preferential attachment principle.

2.4 Network modularity

We studied modularity with six indices: (i) The *VOS Quality index* (*VQ*), was generated by the Pajek layout algorithm that takes into account values (weights) of lines (edges/arcs) as similarities. Similar communities were iteratively drawn closer to each other and the quality index of the final layout with least crossings and closest clusters was given. *VQ* is then calculated as $\sum_{i=1\rightarrow c, j=i+1\rightarrow c} (e_{ij} - a_i^2)$, 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 $1_{v \in ci}$, $1_{w \in cj}$, where m is the sum of weights in the graph and A_{vw} = the weighted value or 0 meaning presence or absence of edge between nodes v and w, in the adjacency matrix A of the network. Finally, a_i is the fraction of weighted k neighbors that are attached to vertices in community i, i.e

$k_i/2m$ (Van Eck and Waltman 2007; Waltman et al. 2010); (ii) The *Clustering Ratio* (*C-ratio*) considers the ratio of the number of node clusters to the count of the connected node set; (iii) The average *Clustering Coefficient* (*C*) describes the mean of the ratio of the triangles to the connected triples for all nodes in the simplified (undirected/no weighted) network (Wasserman and Faust 1994; Ravasz et al. 2002; Barrat et al. 2004). *C* is only and strictly meaningful for unipartite graphs (Newman et al. 2001). We report coefficients of linear regression over *C* for FSF and EFL network projections; (iv) The *Fast Greedy Community* (*FGC*) hierarchical agglomeration algorithm detects community structure with linear running time $O(m\ d\ \log n) \sim O(n\ \log^2 n)$, with m edges, n nodes, and d, the depth of the dendrogram describing the community structure (Clauset et al. 2004); and (v and vi) The *Newman-Girvan* algorithm index (*NG*), computed with default partitions defined by age (NG_{age}) and VOS clustering (NG_{vos}). *NG* calculates the modularity of a network with respect to some division (partition) and measures how good the division is in separating the different node types from each other, to indicate assortative (positive) or disassortative (negative) mixing across modules (Newman and Girvan 2004). *NG* equals $1/(2m)\sum_{ij}(A_{ij}-1/(2m)k_ik_j*\Delta(c_i,c_j))$, where m is the total weights in the graph, A_{ij} are weighted entries in the adjacency matrix, k_i, k_j and c_i, c_j are respectively the weighted degrees and the components (numeric partitions) of nodes i and j, and finally, $\Delta(x,y)$ is 1 if x=y and 0 otherwise (Newman and Girvan 2004). We also computed *NG* for two additional types of membership, *FGC* and *Walk Trap Community* (*WTC*) detection algorithm. *WTC* is similar to *FGC* but computes communities using random walks (Pons and Latapy 2006). *VQ*, *C-ratio*, *C* and *FGC* range from 0 to 1, while the *NG* indices range from -1 to 1. Higher indices represent strong network modularity at a particular event. Heatmaps were customized from scaled modularity matrices with elements given as $(A_{ij}-k_ik_j/(2m))M_{nd}$, where A_{ij}, k_i, k_j and m are as

defined for *NG* (Newman and Girvan 2004) and M_{nd} is network's modularity index at event *nd*. Dendrograms were calculated from 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).

2.4.1 Modularity indices. Modularity indices were calculated for each network using various capabilities of the *igraph* package (Csardi and Nepusz 2006) (BOX 5). Isolated vertices were deleted using a custom function and corresponding partitions were adjusted, as required by the modularity algorithms. We borrowed the single arc of the FSF network at *nd* ~ 0.0045 to make the network at *nd* = 0 non-empty. *cwdn* was used as input in calculation of all modularity indices. VOS modularity indices and partitions were generated using Pajek (Van Eck and Waltman 2007; Waltman et al. 2010). *NG* modularity indices (Newman and Girvan 2004) were computed using two types of memberships (or partitions) as input: VOS clustering (NG_{vos}) and age (NG_{age}). Clustering ratios were determined using custom functions by dividing the number of clusters from the connected node set with the size of the connected node set.

BOX 2.5. R's *igraph* package operations used in modularity analysis

igraph's function	Result
'`read.graph`()'	Import network graph data.
'`fastgreedy.community`()'	Calculate the *FGC* modularity index (Clauset et al. 2004).
'`as.undirected`()'	Collapse directed edges for *FGC*.
'`modularity`()'	Compute *NG* modularity indices.
'`transitivity`()' with the '`average`' option	Determine average clustering coefficient (*C*) (Wasserman and Faust 1994; Ravasz et al. 2002; Barrat et al. 2004).

2.4.2 Heatmaps and dendrograms. *NG* pairwise modularity matrix was computed for each network using '`mod.matrix` ()' (Newman and Girvan 2004). Four types of memberships (or partitions) were used as input: VOS, age, *FGC* and Walk Trap Community detection (*WTC*). *WTC* is similar to *FGC* but computes communities using random walks (Pons and Latapy 2006). *FGC* and *WTC* partitions were extracted by '`membership` ()'. Each matrix was scaled by the absolute value of the overall modularity score of the network, before being drawn as a heatmap. Pairwise scores were adjusted to -1 to make the range [-1, 1]. The x-y axis node labels were color-coded and ordered by age. Dendrograms were generated using '`dist` ()' by calculating squared Euclidean distance matrices that indicate dissimilarities between the cluster means (Borg and Groenen 2003). The distance (or dissimilarity) matrices were hierarchically clustered using '`hclust` ()' with the Ward's minimum variance method aiming at finding compact, spherical clusters (Murtagh and Legendre 2014). Nodes that were clustered in the dendrograms were ordered within clusters by age and were color-coded.

CHAPTER 3: RESULTS AND DISCUSSION

3.1 Tracing the origin and evolution of molecular functions in loops of protein domains

We mapped the evolutionary age of FSF domain structures onto a bipartite graph-theoretic representation of FSF domains and associated EFL prototypes, which we then decomposed into its dual network projections using mathematical properties of finite graphs (Chung et al. 1983). The strategy is described in Figure 3.1A. Since the bipartite graph and its projections describe how prior loop forms that carried primordial functions rearrange to unfold modern repertoires of domain structures and their associated molecular functions, the bipartite network portrays the makeup and evolution of 'elementary functionomes' (EFs). The EF bipartite graph consists of two disjoint sets of entities (nodes), one describing the EFLs (circles) and the other the FSFs (rhomboids). EFL prototypes represent ancestral sequence motifs, 25-30 amino acid residues in length, which do not exist contemporarily but are represented by descendant EFL sequences in modern proteomes (Berezovsky et al. 2000). These sequences that diverged from corresponding prototypes make up loop regions that host important functional roles. In turn, FSFs are considered bona fide structural, functional and evolutionary units of proteins (Murzin et al. 1995). A connection (link or undirected edge) between an EFL and an FSF in the EF bipartite network represents the embedding of that EFL in the structural scaffold of the corresponding FSF domain module. Therefore, network connections describe how the elementary alphabet of functions associates with the structural alphabet of domains at an evolutionary level.

Bipartite graphs project connectivity within each set of nodes (Diestel 2010). Connections between nodes of the two sets become the basis of connectivity between nodes in separate

networks of each set of nodes. The larger the connectivity of nodes of the bimodal graph, the stronger the connectivity in the extracted single node (uni-modal) graphs. In our case, the EF bipartite network **y**ields FSF and EFL networks when projected into its two uni-modal network representations (Figure 3.1A). Links in the FSF network are established when FSFs embody mutual EFL(s) in their makeup. Links of the EFL network arise when EFLs combine to form active sites or are present in separate instances of the same FSF. Thus, the networks describe combinatorial (syntactic) relationships between EFLs and FSFs in which context-dependent rules (pragmatics) determine contextual rules (semantics).

The EF bipartite network *per se* and its FSF and EFL graph projections can only display evolutionary information if an age can be assigned to its node and link components. In other words, mapping age onto the graph representations describes how the networks grow in time (Figure 3.1B). We used a structural phylogenomic analysis based on a census of FSFs in 749 genomes spanning all three domains of cellular life to define a timeline of structural innovation (Wang et al. 2011). The relative age (*nd*) of each FSF domain structure was extracted from this timeline. Calibration points of FSFs associated with microfossil, fossil and biogeochemical evidence, biomarkers, and first-appearance of clade-specific domains (integrated with molecular, physiological, paleontological and geochemical data) converted the timeline of relative ages into geological time scales (Wang et al. 2011). Since prior elementary forms manifest into modern protein functions only when embedded into the structural scaffolds of modern domains, the ages of first-appearance of EFLs were directly transferred from the ages of first-appearance of associated FSFs or second oldest FSFs in multidomain proteins. Other plausible options however produced results similar to those here described.

3.2 Capturing the early history of modern functionomes with time event 'waterfall' networks

The operation of a network system can be modeled as a discrete sequence of events (network growth at temporal intervals) using Discrete Event Simulation (DES) tools (MacDougall 1987; Delaney and Vaccari 1989; Pidd 2004). DES is a form of computer-based modeling widely used to study complex behavior and how interactions between entities are affected by consecutive events. Under DES, the system does not change between events. Consequently, time flows from event to event in discretized manner as a *step function*. Here we borrow the DES rationale to study how discrete evolutionary 'time steps' (intervals) unfold in the growing structure of the undirected EF bimodal network and its uni-modal projections (Figure 3.1B). Events manifest time steps identified by the first appearance of FSF and EFL variants and their mutual links as these create novel molecular functions. We focus on the 43 most abundant profiles defining EFL prototypes out of 138 identified in the 68 archaeal proteomes analyzed (Goncearenco and Berezovsky 2010). We excluded 5 EFLs that were functionally connected to those FSFs without age assignments. The bipartite network of curated 38 EFLs and 82 FSFs resulted in a disconnected undirected graph with few small intra-connected components isolated from a large connected one. The emergent EF network had 134 edges with a *network density* (actual/possible number of edges) of 0.043 [134 / (82×38)] and a *node average degree* (edges per node) of 2.23, i.e. EF groups had ~2 mutual connections on average. The network was well structured. The Visualization of Similarity (VOS) clustering method revealed 25 communities (also known as modules) with a high *modularity index* of 0.894, calculated as standard modularity measure (Van Eck and Waltman 2007; Waltman et al. 2010). The event dynamics of the EF network was made evident by color coding FSF and EFL nodes and arranging them by age in a top-down bimodal

layout following the evolutionary timeline of protein domain structures (Figure 3.2A). The size of nodes was made proportional to the connectivity of nodes, measured by weighted degree, making hub-like behavior explicit in the network structure. In order to better visualize evolutionary patterns, network clusters (comprising of hubs and their constellations) were manually dissected by expanding the horizontal arrangement of the bimodal EF network with the energy-optimized Kamada-Kawai 'separate components' method (Figure 3.2B). The resulting 'waterfall' layout makes evident the processes of functional recruitment as one travels down the events of the waterfall.

The projections of the EF graph were also visualized as waterfall networks (Figure 3.3). To make evolution explicit, uni-modal network connectivity was made directional by connecting nodes with *arcs*, arrows symbolizing the flow of time from older to younger nodes. The EFL and FSF directed graphs were also disconnected and had 113 and 376 arcs, network densities of 0.080 [113 / (38×(38−1))] and 0.056 [376 / (82×(82−1))], and node average degrees of 3.45 and 5.08, implying ~3 FSFs or ~5 EFLs were shared on average, respectively. The uni-modal networks showed significant community structure. The EFL network had 13 clusters with a modularity index of 0.608. The FSF network had a comparable number of 15 clusters but with a higher modularity index of 0.886. The number of outward (outdegree) and inward (indegree) connecting arcs of nodes endowed FSFs and EFLs as 'donors' (sources) or 'acceptors' (sinks) of functional loops and domains, respectively. The horizontal and vertical scaling of nodes were made proportional to their weighted outdegree and indegree, respectively. This facilitated the visualization of hubs. The transition from wide to tall symbols along the flow of time revealed

the source-sink origination dynamics of molecular functions (see below). This transition expresses the expected increase in probability of co-opting older EFLs and FSFs with time.

Two major waves of functional innovation were evident in the waterfall diagrams of the EF network and its projections (Figures 3.2 and 3.3). These waves had separate origins and involved sandwich, barrel and bundle protein domain structures. Wave 1 was the larger of the two and originated in the P-loop containing nucleoside triphosphate (NTP) hydrolase FSF (c.37.1) and it's uniquely connected and relatively long *p*-loop EFLs 7, 6488 and 6739. The c.37.1 FSF is a Rossmanoid α/β/α-layered domain structure that is the most ancient and popular in the timeline of domain history (Caetano-Anollés et al. 2009a; Caetano-Anollés and Caetano-Anollés 2013). The *p*-loop EFLs crucially enabled the nucleotide triphosphate binding functions of the P-loop hydrolase fold with its Walker A (*p*-loop) sequence motif located at the elbow, usually connecting the first β-strand of the main β-hairpin loop (or loop derivatives) that binds to di- and tri-nucleotides. The EFL network shows that the '*p*-loop' wave established several massive pathways of EFL recruitment involving four cysteine-rich EFLs, the EFL 536 hub, the downstream highly connected EFLs 1845 and 1632, and terminal EFL 2524 (Figure 3.3A). These cysteine-rich EFLs of the wave involved a strong recruitment pathway spanning ~0.5 Gy of history, in which the NAD(P)-binding Rossmann-fold FSF (c.2.1) and the S-adenosyl-L-methionine-dependent methyltransferase FSF (c.66.1) with their 3-layered α/β/α structures, and the OB-fold of the nucleic acid-binding protein FSF (b.40.4) with its closed or partly-opened β-barrel structure, enabled a host of functions related to metabolism and translation. In particular, the cysteine-rich metal binding loop of EFL 1845 formed a cysteine nest that coordinated Zn^{2+} metal binding necessary for interactions with nucleic acids in 13 EFL-related FSFs. Among these

FSFs were the ancient OB-fold structure of b.40.4, class II aminoacyl-tRNA synthetases and biotin synthetases (d.104.1) and nucleotidyltransferase (d.218.1) FSFs with α/β/α-layered and sheet structures, and more derived beta and beta-prime subunits of DNA dependent RNA polymerase (e.29.1), RNA polymerase subunit (g.41.9) and prokaryotic type I DNA topoisomerase (e.10.1) FSFs with β-barrel and winged helix-like structures (Figure 3.2B). Finally, the EFL 536 hub also linked downstream glycine and glutamate-rich EFLs 3314 and 7009 and the glycine-rich nucleotide-phosphate binding EFL 8 that is typically embedded in β/α-barrel structures widespread in metabolism via the Rossmann c.2.1 structure. EFL 8 acted as hub for other downstream EFLs, including EFLs 7009 and 3619.

The second wave originated in the 'winged helix' DNA binding domain FSF (a.4.5) and its uniquely connected EFL 2914. The wave appeared soon after the *p*-loop wave but was much constrained in scope; part of it merged with the *p*-loop wave through EFL 3619. The a.4.5 FSF harbors the DNA/RNA-binding 3-helical bundle fold (a.4) structure, which is flanked by a 4-strand β-sheet. The structure exposes crucial elbows between the helix-turn-helix motifs that harbor the specificity of protein-protein and protein-RNA interactions typical of these enzymes. The winged domain structure plays central roles in transcription, providing flexibility and nucleic acid clamping capacity to RNA polymerases (Teichmann et al. 2012). The structure also provides crucial surfaces for domain-domain recognition in complexes (like polymerases, ubiquitin-ligases, condensins) and other protein-protein interactions.

It is remarkable that the first two waves uncovered by the EF network and its projections involve the same primordial sandwich α/β/α-layered structures, β-barrels and helical bundle structures we

identified as part of the first 54 domain families that appeared in evolution (Caetano-Anollés et al. 2012). It is also remarkable that the '*p*-loop' and 'winged helix' waves embedded the first two major gateways of enzymatic recruitment we previously identified in metabolism, the first gateway mediated by the c.37 fold and originating in the energy interconversion pathways of the purine metabolism subnetwork, and the second mediated by the a.4 fold and originating in the porphyrin and chlorophyll metabolism subnetwork and the biosynthesis of cofactors (Caetano-Anollés et al. 2007; 2009b; Caetano-Anollés and Caetano-Anollés 2013). The fact that we are obtaining congruent evolutionary results with different data sets supports the historical statements we here propose.

3.3 The evolutionary dynamics of emergence of molecular functions

The waterfall layout unfolded 61 unique events in the EF and FSF networks and 26 events in the EFL network along a timeline that spans the origin of proteins (*nd* = 0) and the present (*nd* =1) (Figure 3.2). The distribution and connectivity of EFL and FSF nodes within and across these events provides information about how dynamic, recurrent and widespread is the combinatorial recruitment process that embeds loops into domains scaffolds and generates new molecular functions. Since the EFLs that were studied are the most abundant in genomes, they are likely the oldest (Goncearenco and Berezovsky 2010). Indeed, the largest hubs bolstering most of the connectivity of the networks of EFLs and FSFs we sampled unfolded very early in protein evolution (Figures 3.2 and 3.3). EF innovation unfolded during the first ~1.8 Gy of protein history (Figure 3.2). However, the combinatorial recruitment process involved the entire timeline; most acceptors of EFLs and FSFs populated the *nd* = 0.0-0.1 and *nd* = 0.1-0.5 ranges, respectively. These patterns can be made quantitative by dissecting the source-sink relationships

and evolutionary span of network connectivity with bar plots describing the chronological accumulation of links along the timeline (Figure 3.4). Further insight can be obtained from box-and-whisker plots of accumulation of weighted indegree and outdegree (Figure 3.5) and patterns of contraction or expansion of mutually-facilitated EFL and FSF innovation, extracted from the distributions of total degree (Figure 3.6). Overwhelmingly, sink EFLs acted as acceptors of very ancient EFLs of $nd < 0.1$. In contrast, sink FSFs drew innovation from domains spanning the entire timeline, taking at the same time advantage of the repertoire of very ancient EFLs. Individual FSFs however co-opted a significant number of ancient domains for their functional tasks, confirming evolutionary patterns of recruitment obtained in the enzymatic analysis of metabolic networks (Kim et al. 2013).

Connectivity patterns make explicit the evolutionary dynamics of emergence of molecular functions, falsifying some alternative hypotheses that could explain it. Historically, the creation of molecular novelty most likely involved the ligation of dipeptides and small polypeptides with limited ordered structure, followed by the formation of larger peptides harboring stable loop structures (Caetano-Anollés et al. 2012; 2013), and finally the combination of loops to form defined 3D folded topologies in small protein domains (Berezovsky and Trifonov 2001; Aharonovsky and Trifonov 2005; Goncearenco and Berezovsky 2010). While a continuum of these prior forms is expected when invoking the principle of spatiotemporal continuity that supports evolution, Figure 3.4 reveals that the relative formation of useful loops and domains in the two waves of recruitment and innovation occurred (and is occurring) at different rates. A quick and early discovery of loops provided the raw materials for their combination in domains along the entire span of protein history. However, the generation of novel EFLs and FSFs

appears ongoing and their use in combination with older domains suggests that old EFLs are still evolutionarily active; they are not relics tagged for extinction but evolvable forms. Thus, the fast establishment of highly conserved sequence motifs in elbow regions of loops and their combinatorial use serve to define hierarchical levels of structural complexity, which appear tightly interrelated throughout protein history. Remarkably, phylogenomic studies have also shown this same kind of dynamics materializing with domains and their combinatorial use in multidomain proteins (Wang and Caetano-Anollés 2009). It is noteworthy that connectivity patterns falsify evolutionary scenarios of sequential but separate build-up of EFL and FSF repertoires. Both EFLs and FSFs unfold together in evolution, but at different rates. Despite their high evolutionary conservation, results also falsify the possibility that EFLs are 'molecularly canalized' forms that resist evolutionary change. The EFLs prototypes arise from diverse families of sequence motifs, which express themselves in different protein structural contexts. Finally and from a historical point of view, only functionally useful prior forms would have prevailed if they provided properties that would extend the persistence of the emerging cells, including membrane stability and transport modulation, bioenergetics, and peptide-cofactor biosynthetic functions.

3.4 Emergence of preferential attachment behavior typical of scale-free networks

Networks whose dynamics follow the preferential attachment principle harbor large, highly connected hubs that attract increasingly more links in a 'rich-get-richer' fashion. In these highly inhomogeneous networks, which are remarkably popular in biology, the probability $P(k)$ of a node being linked to k other nodes (i.e. the fraction of nodes with k links or k-neighbors) decays as a power law, $P(k) \sim k^{-\gamma}$, without a characteristic scale. 'Scale-free' networks of this kind

generally have exponents γ = 2.1–2.4, driving a heavy-tailed distribution in which very few nodes have high connectivity degrees (Strogatz 2001). For metabolic reaction networks γ = 2.2 in all organisms (e.g. Jeong et al. 2000). In order to test if the evolving EF network and its projections had a tendency to follow the scale-free distribution, we studied the chronological accumulation of connections (links or arcs) of the growing networks and tested power law behavior with appropriate statistics (Figures 3.7A, 3.8 and 3.9). Remarkably, we found that the power law and associated generative models are an emergent property of the EF network but not of its projections.

A number of statistics failed to reject power law behavior in very ancient connections of EFLs and most connections of FSFs unfolding in the EF bipartite network (Figures 3.7A and 3.8). Thus, co-options of ancient EFL by multiple FSFs and FSFs of all ages by multiple EFLs follow scale-free properties (Figure 3.7A). The most prominent indicator of rejection was the Kolmogorov-Smirnov (KS) statistical test of power law fit (Newman 2005; Clauset et al. 2009). Low p-values of the KS test (<0.05) and high values of the KS fit statistic (>0.10) rejected 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, only increased in degree distributions of FSF nodes of the growing EF networks. These patterns are indicative of power-law decay. Finally, the log-likelihoods of the fitted power law parameters were relatively much lower than zero for most networks of the timeline, making power law distribution less likely and discontinuous in any of these networks.

The statistical analyses of the timeline of growing EF networks therefore reveal a surprising property of power law emergence. The early-evolved EFL 'prior form' component of the bipartite network transfers power law behavior to the FSF domain component as molecular functions unfold in protein evolution (Figures 3.7A and 3.8). *Log* based linear regression models overlapping with power law curves show that the coefficient of power law decay γ for the FSF portion of the EF network increases with time and reaches a limit of 1.8, which is somehow lower than γ reported for metabolic networks (Ravasz et al. 2002). For the EFL portion, however, γ starts with ~2, but then quickly plummets to ~1 quite early in protein evolution (*nd* ~ 0.2). The coefficient of determination (R^2) of ~80% (but never <50%) supports the linear models. Remarkably, EFL and FSF network projections maintain a lower average $\gamma < 1$, consistent with statistics that reject power law behavior of these network projections. The unexpected result of power law transfer from EFL to FSF components may be indicative of a global scaling phenomenon in the biphasic emergence of biological modules (Mittenthal et al. 2012), which we now explain.

3.5 Hierarchical modularity and the rise of primordial functions

Networks are modular when they embed communities (modules) of nodes that connect preferentially to each other within bounds of a community (Newman and Girvan 2004). Modularity counteracts the scale-free property of biological networks by equalizing the degree distribution of nodes in communities (Overbeek et al. 2000; Jeong et al. 2000; Wagner and Fell 2001). However, both properties can be reconciled when modules are integrated hierarchically (Ravasz et al. 2002). Modularity is primarily measured with the *average clustering coefficient* (*C*), the ratio of triangles (graph cycles of length 3) to the connected triples in the graph,

averaged over all nodes, ignoring the direction and weights of the edges (Wasserman and Faust 1994; Ravasz et al. 2002; Barrat et al. 2004). Since bipartite networks have no triangles, we studied the modular organization of the EF network through *C* of its projections (Figures 3.7B and 3.10). The FSF and EFL networks exhibit *C* of ~0.83, much higher than ~0.6 reported for metabolic networks (Wagner and Fell 2001; Ravasz et al. 2002). Elevated *C* of EF network projections suggests various densely intra-connected modules of loop motifs and domain structures integrated by few inter-modular links. The EF network must therefore embody a highly consolidated modular structure.

C of scale-free models sharply declines with network size N as $N^{-0.75}$ (Albert and Barabási 2002), contrary to being independent of N if the networks are highly modular (Ravasz et al. 2002). For the FSF and EFL networks, *C* regressed with N as $N^{0.0022}$ and $N^{-0.006}$, and with age *nd* of the networks as $nd^{0.17}$ and $nd^{-0.15}$ (Figures 3.7B and 3.10), respectively, confirming the modular structure of the evolving networks. As expected, *C* of strictly power-law (Barabási) reference controls were zero (Newman et al. 2001). The timeline of growing FSF and EFL networks unfolded trends of modularity and scale-free behavior that were anticorrelated. For example, the *C* of the FSF network shows an initial decline and then a rise in modularity (starts with 0.8, drops to 0.5 at *nd*~0.063, followed by growth to 0.875. This trend matches the KS fit indegree statistic that eventually rejects power law behavior (starts with 0.3, drops to 0.125 and then rises to 0.375) (Figures 3.7B, 3.8 and 3.10). Since the counteracting trends of modularity and preferential attachment of the FSF and EFL networks must impact the emergent scale-free behavior of the EF bipartite network, our findings lead to a noteworthy conjecture. Transfer of scale-free properties

from EFL to FSF components of the functionome involve the generation of modular and hierarchical structure of interacting loop motifs and domain structures.

To test this conjecture, we analyzed the dynamics of three additional 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 with high measures of inter-community centrality (Newman and Girvan 2004). Removal uncovers the community structure of a network measured by the *NG* index. *NG* partitioned by age (NG_{age}) ranges from –1 to 1. Positive values indicate modular connectivity within events while negative values indicate connectivity across them. *NG* partitioned by VOS (NG_{vos}) describes the cohesiveness of VOS divisions (Van Eck and Waltman 2007; Waltman et al. 2010). Finally, the Fast Greedy Community (*FGC*) detection algorithm provides a hierarchical perspective of agglomerative community structure (Clauset et al. 2004). Remarkably, we found that NG_{vos} and *FGC* indices unfolded similar patterns of growth of community cohesiveness and agglomerative structure with age for all growing networks. Conversely, NG_{age} dissected divergent dynamic behaviors in the EF network and its projections. The EF network unfolded an age-linked modular structure along the timeline, with an initial spike of NG_{age} ~0.5 (*nd*~0) followed by a gradual decrease to ~0.25 (*nd*~0.37). The FSF and EFL projections developed instead an age-independent modular structure, with initial NG_{age} of about –0.5 and –0.25 (*nd*~0), respectively, which quickly flattened towards 0 (*nd*~0.1) (Figures 3.7B and 3.10). In these networks, communities of nodes with various ages are indicative of modularity patterns of recruitment. It thus appears that during functionome emergence, loop motifs and domain structures of the EF network were tightly coupled by age. This trend

diminished as network agglomerative modularity matured and existing forms engaged in widespread recruitment of emerging constructs throughout the timeline. The recruitment trend is 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 (Figure 3.7C, Video 1). Evidently, these patterns were induced by a parallel development of hierarchy that encouraged modules to cluster within modules in a fashion not distinct from the scale-free organization of modules in metabolic networks (Ravasz et al. 2002) (Figure 3.11, Video 2). In contrast, the growth of EFL and FSF networks was initially driven by recruitment, a trend that was quickly but moderately counteracted by age-bound modularity. Thus, the transfer of scale-free properties from loop motifs to domain structures involves a hidden switch to modular and hierarchical structure, which occurred ~3.4 Gy ago. Remarkably, the timing of this switch coincides with the early development of genetic code specificity in emerging aminoacyl-tRNA synthetases and the ribosome (Caetano-Anollés et al. 2013) that was facilitated by the OB-fold structure (Figure 3.2).

CHAPTER 4: CONCLUSIONS

Tracing evolutionary age onto growing EF networks and their projections uncovered two clear waves of functional innovation that involved ancient EFLs and founder '*p*-loop' and 'winged helix' domain structures. We find that the dynamics of recruitment of loop motifs by domain structures is ongoing and highly modular. The emergence of molecular functions showed properties of hierarchical modularity and emergent power law behavior. Modules in the EF network behaved as integrated communities of interacting structural parts defining classes of molecular functions. Our analyses suggest that module emergence occurred through a biphasic process of diversification (Mittenthal et al. 2012). In a first phase, the parts of the emerging functionome (loop motifs and domain structures) associated massively through processes of recruitment. Their linkage was weak. As motifs and structures diversified, they engaged in competition and were selected for functional performance. Useful emerging interactions constrained their associations. This caused tightly linked parts to self-organize into modules expressing as closely-knit communities of interactions. In a second prolonged phase, variants of the modules evolved and now became parts of a new generative cycle of higher-level organization governed by scale-free module recruitment that is still ongoing in biology. The generation of biphasic patterns of change may be a general phenomenon in network biology.

CHAPTER 5: FIGURES

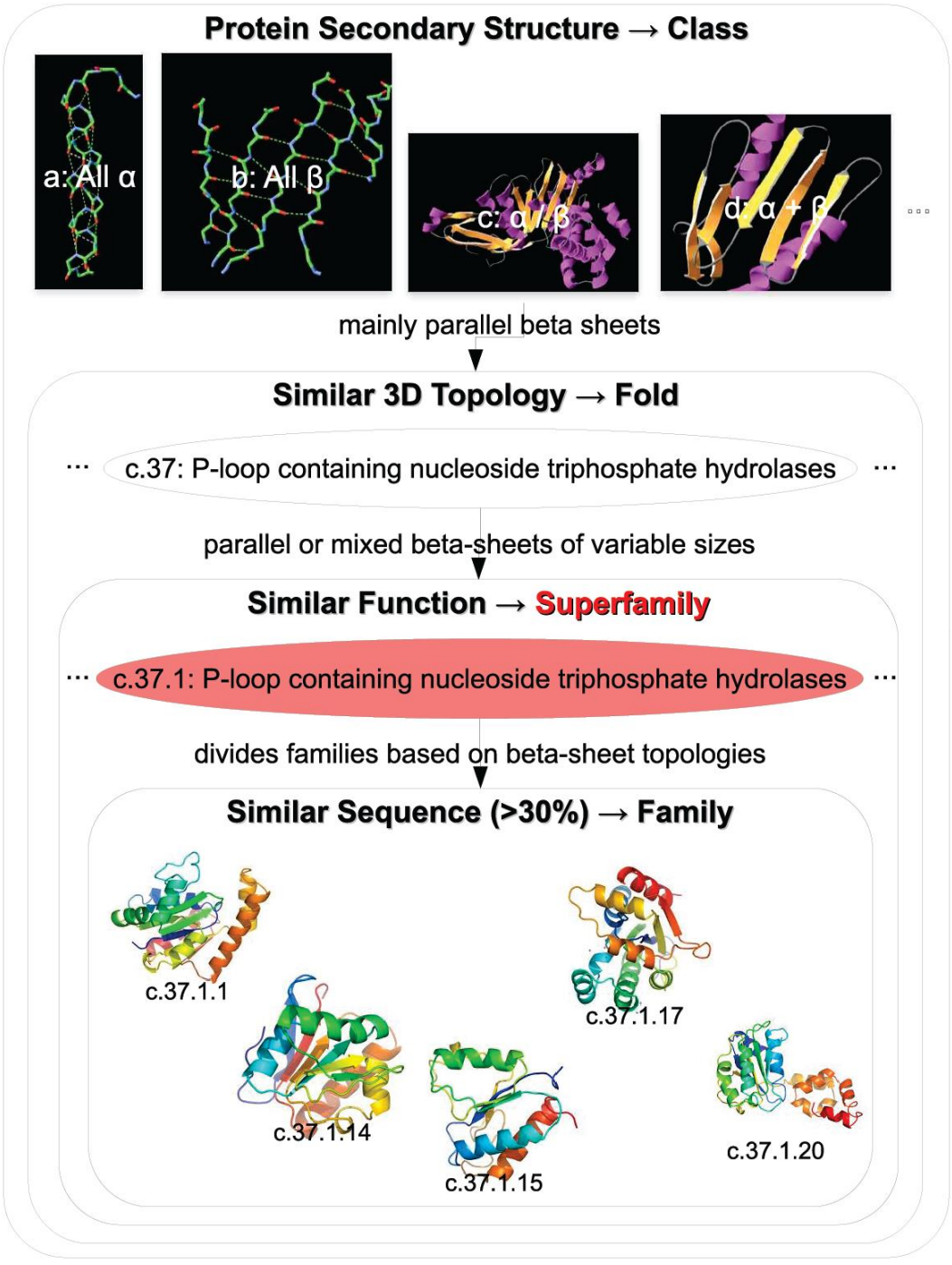

Figure 1.1. SCOP classification scheme. The protein architectures worked upon in this study were from the superfamily level. Image sources: *Protein Secondary Structure Types* [3D Model]. (2012). Retrieved Dec 22, 2013, from: http://www.proteinstructures.com/Structure/Structure/secondary-sructure.html; *Sample c.37.1 families* [3D Model]. (2009). Retrieved Dec 22, 2013, from: http://supfam.mbu.iisc.ernet.in/CGI/display_pali.cgi

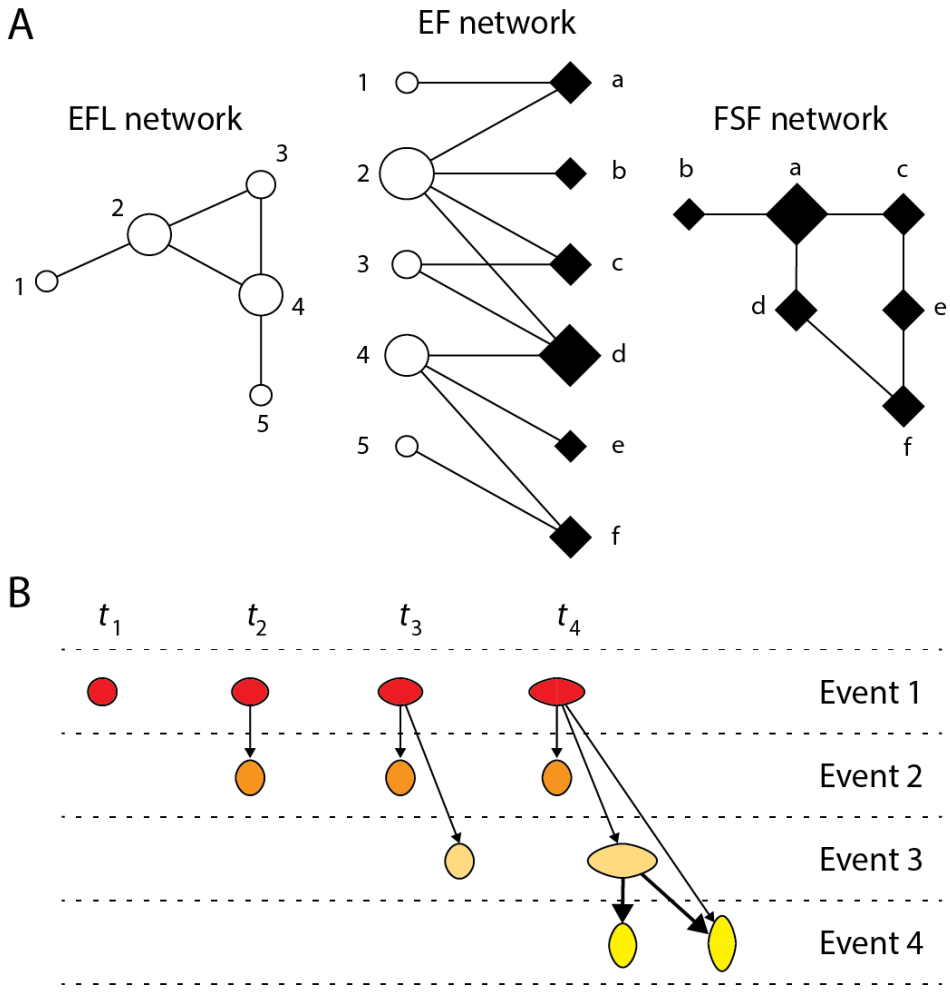

Figure 3.1. Using bipartite networks to study the evolution of elementary functionomes (EFs). (A) The diagrams illustrate an undirected bipartite EF network and its EFL and FSF network projections. Nodes are described as symbols, with size proportional to the number of links they establish. **(B)** Construction of directed 'discrete event' networks in waterfall format. As time progresses from left to right, events describe the progressive appearance of nodes and links. Network growth is made explicit by coloring and arranging nodes according to age (red-to-yellow and top-to-bottom), using time-induced arrows (arcs) with density proportional to their connectivity, and horizontal and vertical sizes of symbols proportional to the respective 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. Note how the chronological accumulation of connections in the originating node (colored red) progressively increases its outdegree as connections are established with nodes of later events. This widens the symbols horizontally. In contrast, nodes appearing later in time receive connectivity from earlier nodes, increasing their indegree and heightening the vertical scale of symbols.

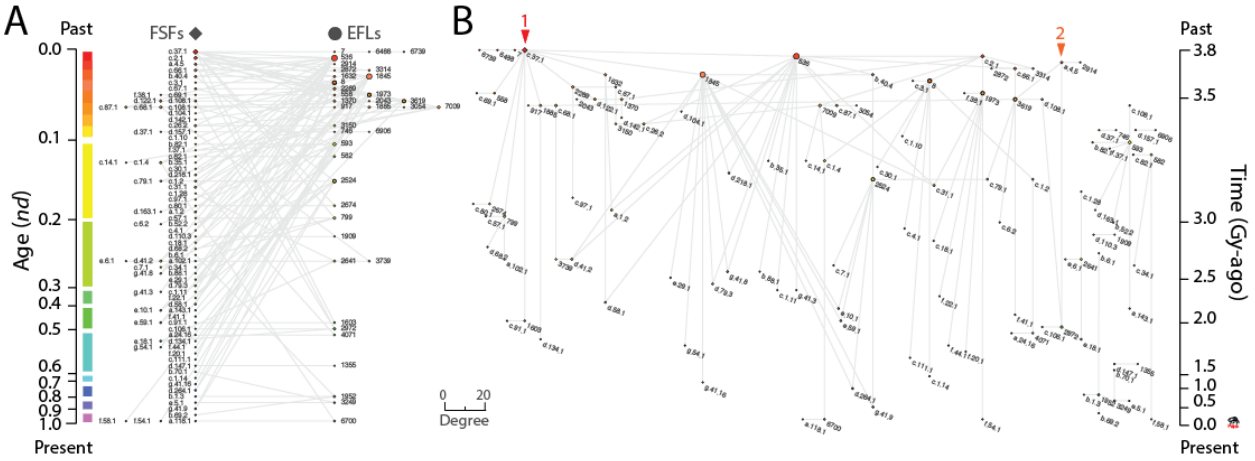

Figure 3.2. The EF network in bipartite (A) and waterfall (B) layouts. EFL and FSF nodes were arranged top-down according to age (*nd*) displayed in a relative 0-to-1 scale, labeled using established SCOP nomenclature (Murzin et al. 1995) and colored according to time events (left). Ages were also time-calibrated with a molecular clock of FSF structures 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. Red arrowheads indicate the origin of major waves of recruitment in the time event waterfall.

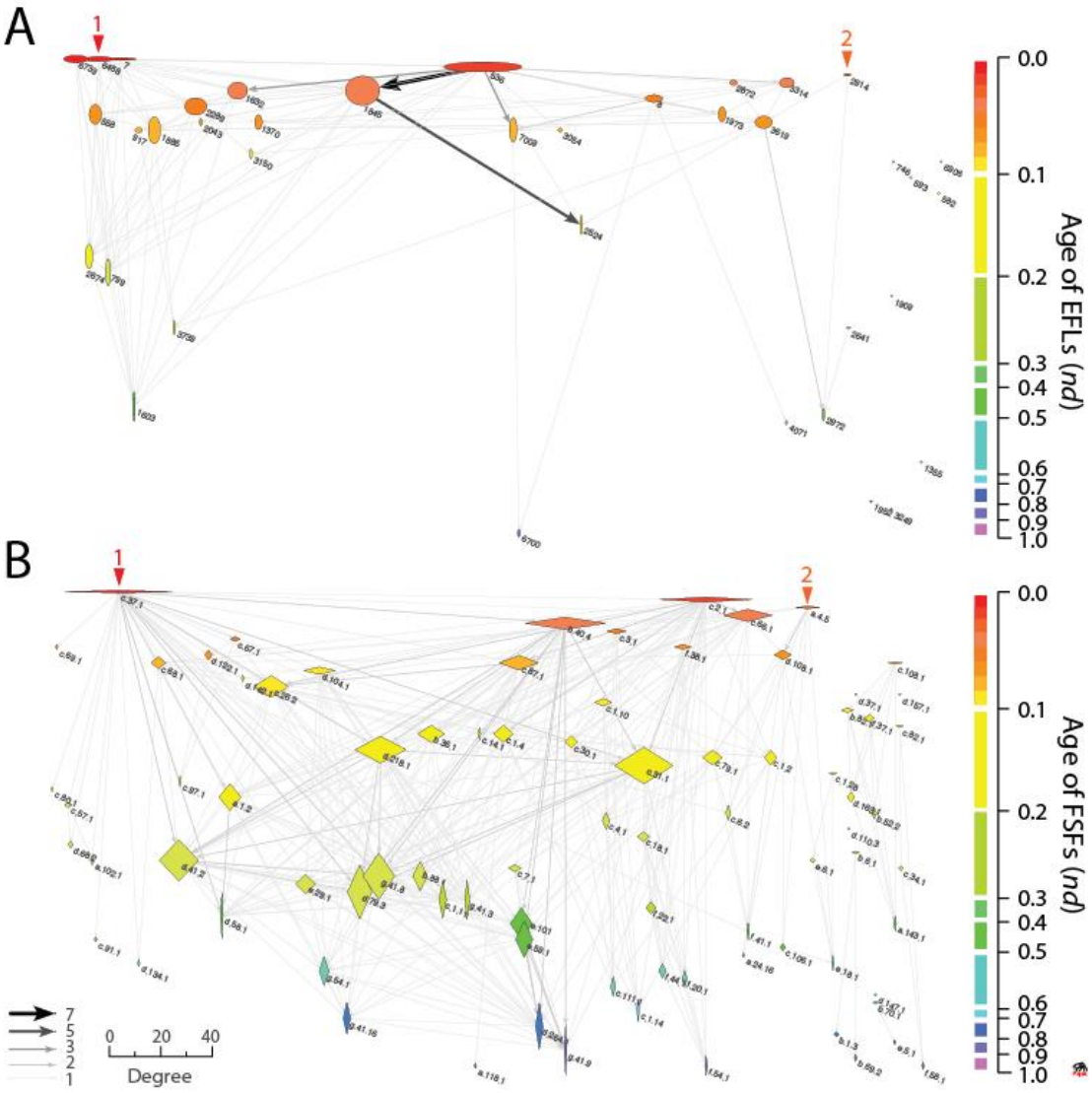

Figure 3.3. EF network projections in waterfall layout. (A) EFL network defined by 38 EFLs (ellipsoids) and arc connections (arrows) representing sharing of FSF structures. **(B)** FSF network defined by 82 FSFs (rhomboids) and arc connections representing sharing of EFL motifs. EFL and FSF nodes in their uni-modal 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). The 2D scale of nodes was kept proportional to their weighted degree. In particular, the horizontal and vertical sizes of the ellipsoids (EFLs) and rhomboids (FSFs) were made proportional to the weighted outdegree and indegree, respectively, showcasing source-and-sink relationships. All weighted degree vectors were scaled by a factor of 10 to avoid vanishing of 0-degree entities. The width of arcs joining the EFLs and FSFs was made proportional to the number of 'shared FSFs and EFLs', respectively. The same criterion stands true for the grey scale of the arcs. The arcs symbolize the flow of time (random direction for contemporary nodes).

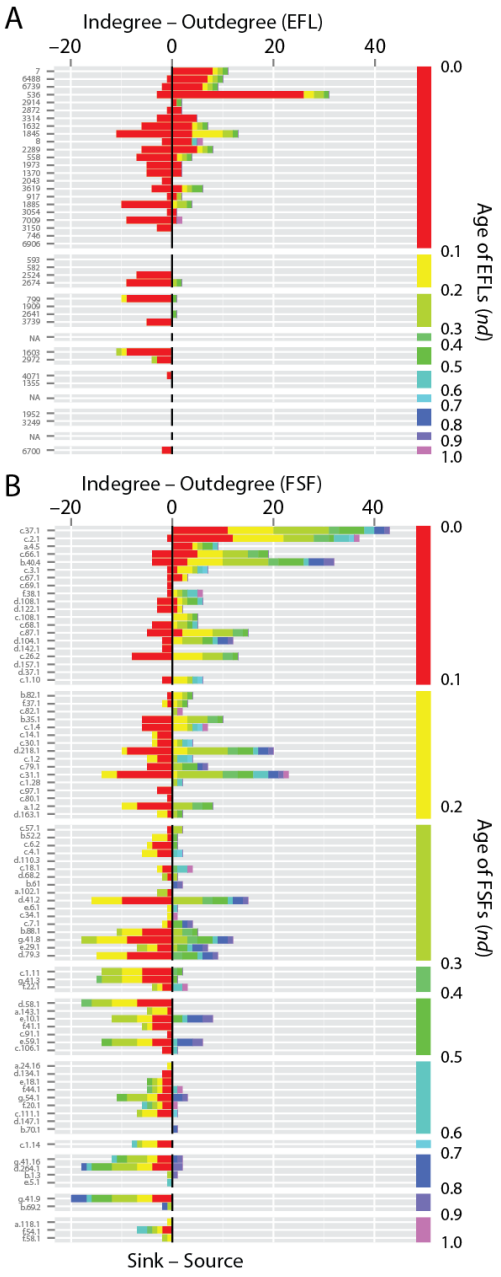

Figure 3.4. Chronological accumulation of connectivity in EFL and FSF networks. The stacked bar charts depict the chronological accumulation of connections (arcs) in events along the timeline of EFL **(A)** and FSF **(B)** innovation, which are labeled using standard SCOP nomenclature (Murzin et al. 1995). Each event corresponds to the discovery of EFLs and FSFs from one of 26 and 61 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-purple). For each node in non-vacant bins, 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 elementary functions viewed from the perspectives of loops and domains.

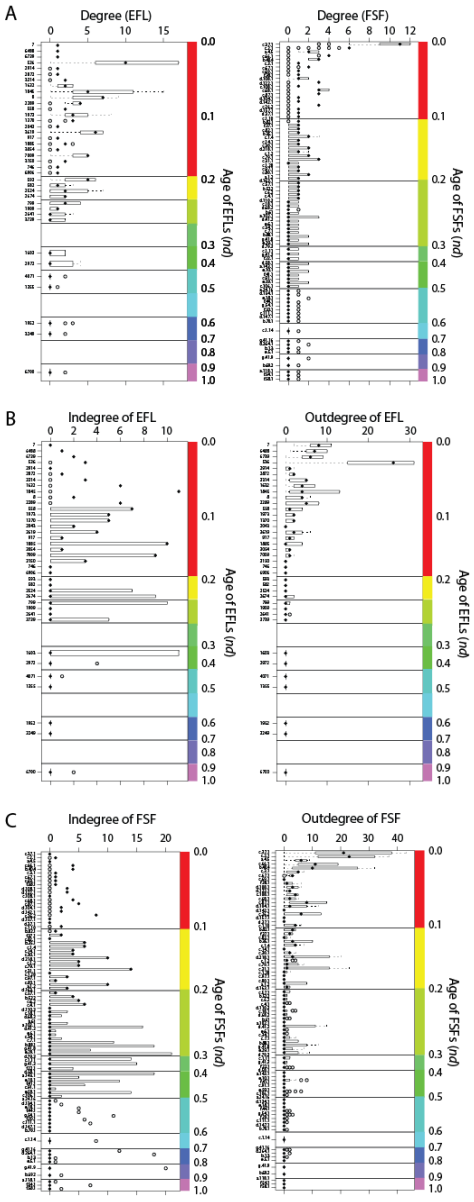

Figure 3.5. Boxplots of the EF bipartite network (A) and its EFL (B) and the FSF (C) network projections. The Tukey boxplots show boxes (first and third quartiles bracketing the median), whiskers (values within ±1.5 × inner quartile range), and outliers. The plots describe the span of connectivity given as expansion in degree of each node along the events of evolutionary timeline. They provide a view of chronological accumulation of connections (in the form of links or arcs) with time. For the EF network, connectivity of EFL and FSF portions is given separately. For the projected EFL and FSF networks, arc connectivity is given as node indegree and outdegree and plotted separately. Each event corresponds to the discovery of EFLs and FSFs, labeled using standard SCOP nomenclature (Murzin et al. 1995), from one of 26 and 61 events, respectively, along a timeline, which are labeled using relative 0-to-1 scale on the right of the plots. The timeline of events was coarse-grained into 10 age bins (colored red-to-purple).

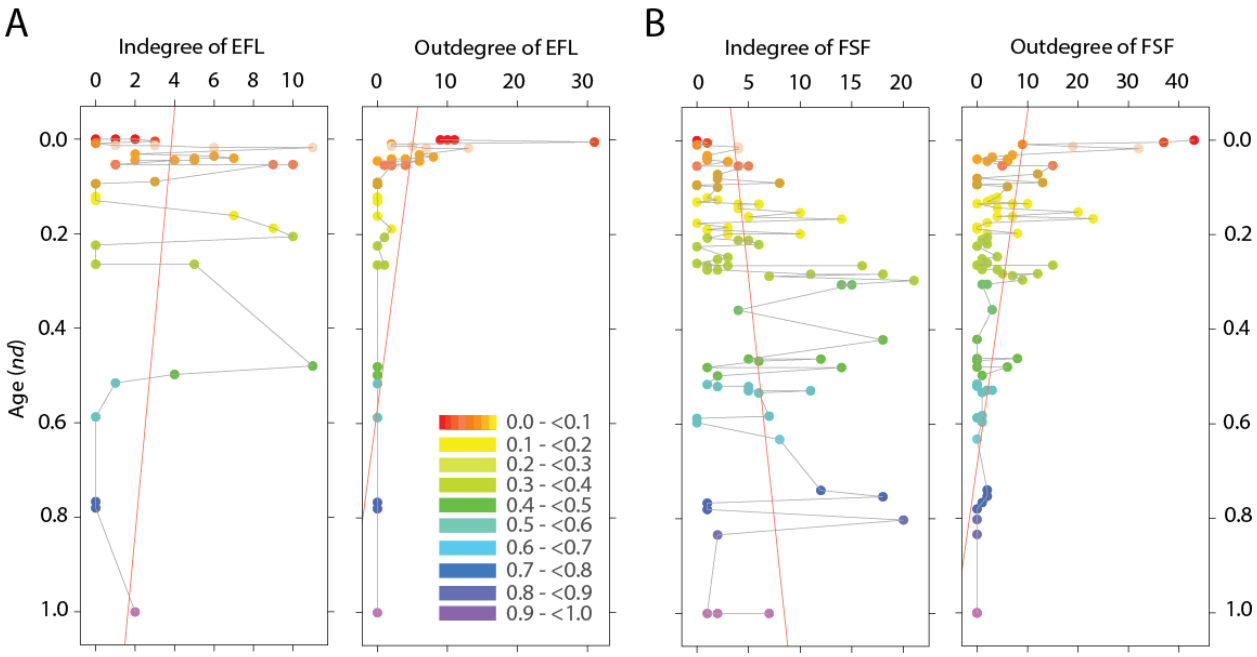

Figure 3.6. Connectivity of events along the timeline in the projections of the EF network, the EFL (A) and the FSF (B) networks. Data points describe indegree and outdegree of nodes of different ages of networks analyzed at present time (*nd* = 1). Symbols were color-coded according to node age. Linear regression lines (red) show that the only trend of growth of connectivity with age is the embedding of old EFLs in emergent FSFs to make new molecular functions (indegree of panel B).

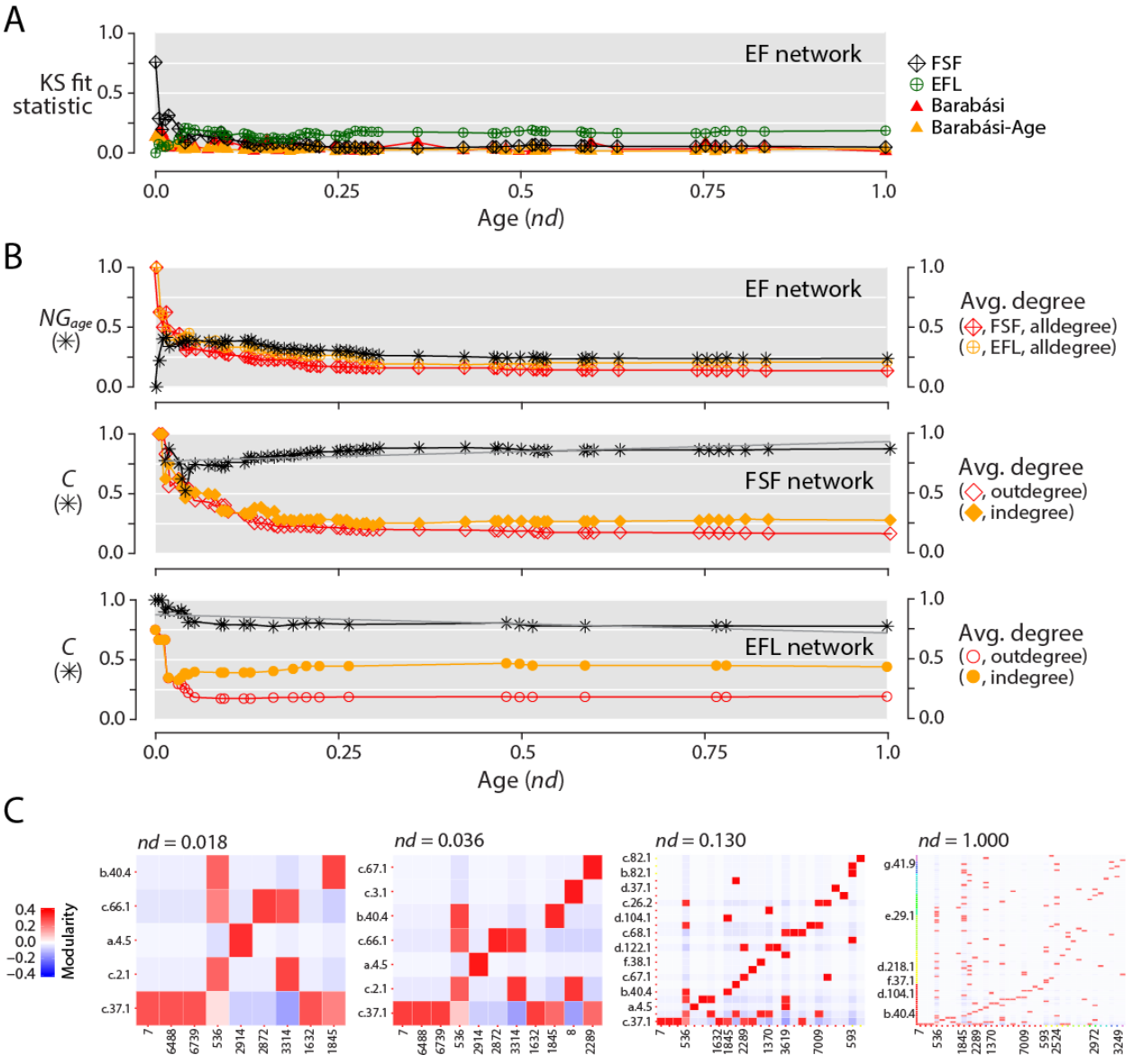

Figure 3.7. Scale-free and modular network behavior. (A) Transference of the scale-free property in the EF network. The KS fit statistic measures network-degree deviations from the fitted power law distribution. Lower KS indicates better fit (Newman 2005; Clauset et al. 2009). The reference Barabási (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 EF 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 FSF and EFL 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 (grey) over *C* for the FSF network as 0.0022 by network size (*N*) and 0.17 by age, and those for the EFL network as -0.006 by *N* and –0.15 by age. Linear regression lines are shown only by age. 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 EFL and FSF portions of the EF network and for the 'outdegree' and 'indegree' of EFL and FSF networks. 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 EF network. The cells of the heatmaps represent modular strength between an EFL and FSF, as compared to their individual connectivity with the rest of the network, scaled by 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 EFLs and FSFs. The last panel corresponds to the fully-grown EF network. The significant loop motifs and domain structures involved in the two major waves of functional innovation are displayed using established SCOP nomenclature (Murzin et al. 1995), ordered ascendingly and color-coded according to node age.

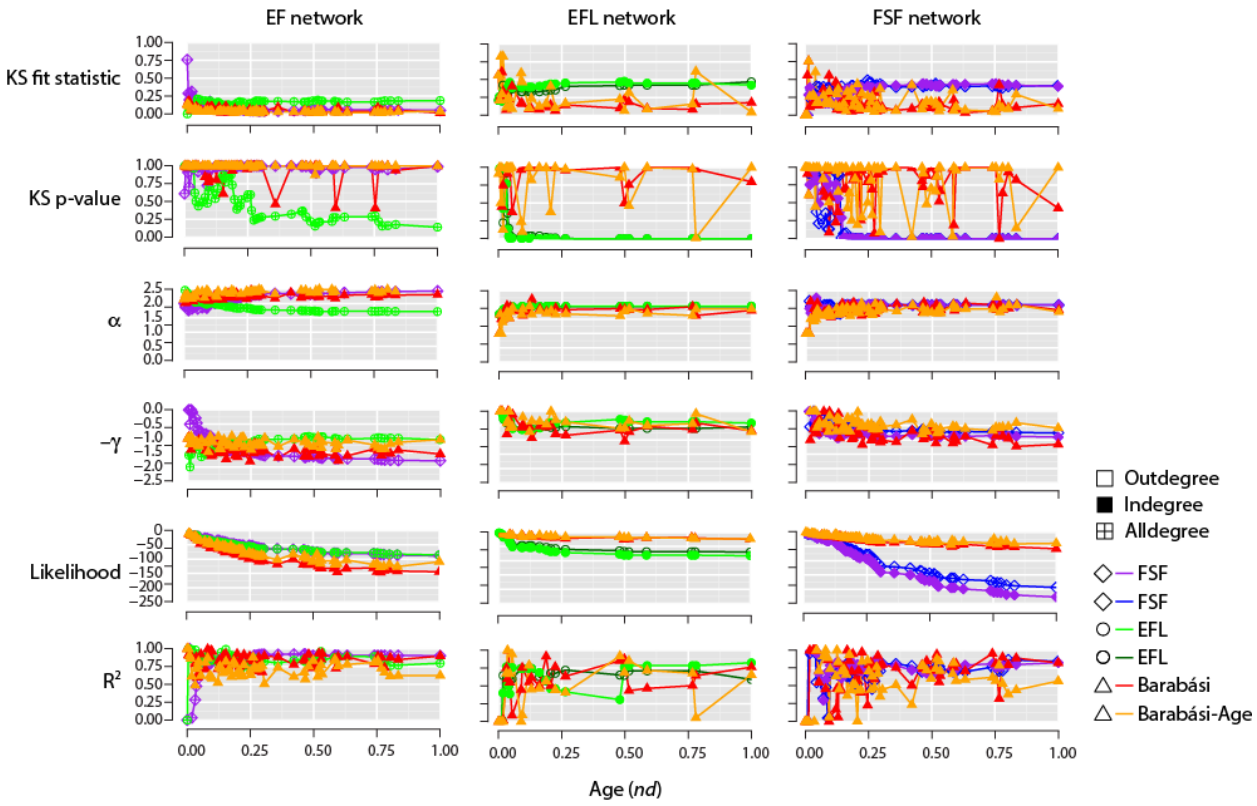

Figure 3.8. Statistical descriptors of power law behavior for the EF bipartite network and its EFL and FSF network projections. Six differential indicators of preferential attachment were plotted for each network. Two reference curves, Barabási (red) and Barabási-Age (orange), were included for comparison. Barabási specifies the probability of preference of an old node as $P_i \sim k_i^{\alpha}$ while Barabási-Age confers heavier power law properties to older (with smaller *nd*) nodes using $P_i \sim (k_i^{\alpha})(l_i^{\beta})$. Here, k_i is the indegree of node i in the current event; α is the preferential attachment exponent, with default value one, i.e. 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; β is the exponent of aging, kept at 1 for linear increases in probability of preference of an older node (with high l_i). Separate curves were computed for the 'alldegree' of EFL and FSF portions of the EF network and for 'outdegree' and 'indegree' of projected EFL and FSF networks. Degrees were cumulative and weighted. The six power law indices include: (i) the KS fit statistic that scores deviation of input degree data vector from the fitted power law distribution; smaller scores denote better fit; (ii) the KS p-value less than α=0.05 indicate rejection of the null hypothesis of degree data being drawn from the fitted power-law distribution (Newman 2005; Clauset et al. 2009); (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 (likelihood); and the (vi) coefficient of determination (R^2) indicating confidence in the linear model through percentage of degree data explained by the regression line. Higher α, lower -γ and closer to zero likelihood correspond to power law behavior. Statistics were calculated for each event of the growing networks. Age (*nd*) is indicated on a relative 0-to-1 scale.

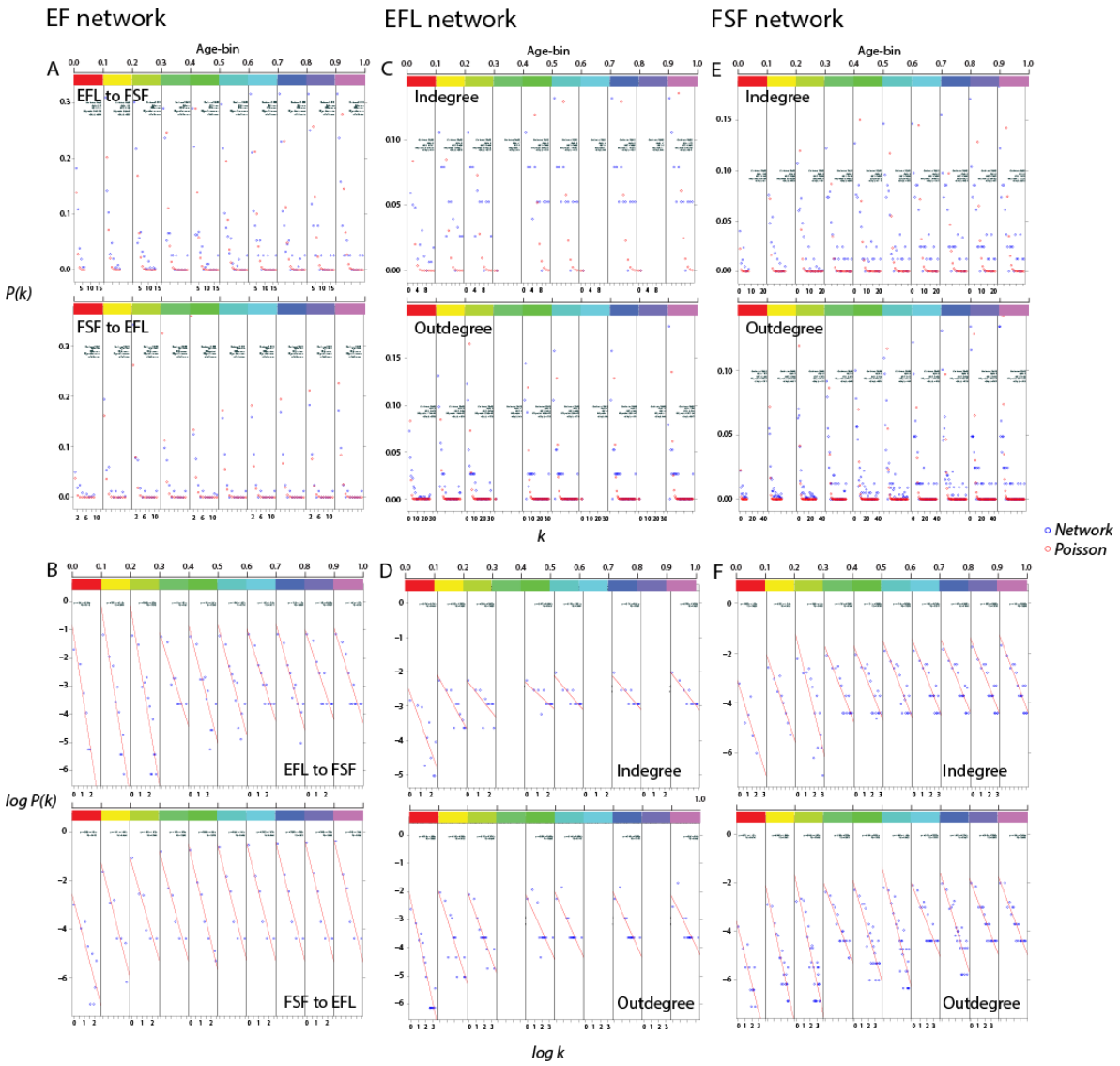

Figure 3.9. *P(k)* vs. *k* and *log-log* plots of the EF bipartite network (A,B) and its EFL (C,D) and FSF (E,F) network projections. The *P(k)* vs. *k* graphs (blue) describe the relation of the range of cumulative weighted degree (1 – maximum), *k*, and the probability of occurrence of degree values, *P(k)*. The *P(k)* vs. *k* plots include a control *Poisson* distribution (red) and describe four power-law statistical descriptors: KS fit statistic, KS p-value, α and likelihood (-2 log L) (Newman 2005; Clauset et al. 2009). *Log-log* graphs show a scatter-plot of natural log of *k* against *P(k)* (blue). A linear regression line (red) over the *log-log* data points is included and the slope -γ and coefficient of determination R^2 of the linear model are displayed in the inset. For the EF network, separate plots were generated for the EFL and FSF portions. For the projected FSF and EFL networks, indegree and outdegree *P(k)* vs. *k* and *log-log* graphs were plotted separately. Cumulative weighted degree was coarse-grained into 10 age-bins (colored red-to-purple), corresponding to the discovery of EFLs and FSFs from one of 26 and 61 events, respectively. Notice that few age-bins in EFL network are empty. The timeline is displayed in a relative 0-to-1 scale on the top of the plots.

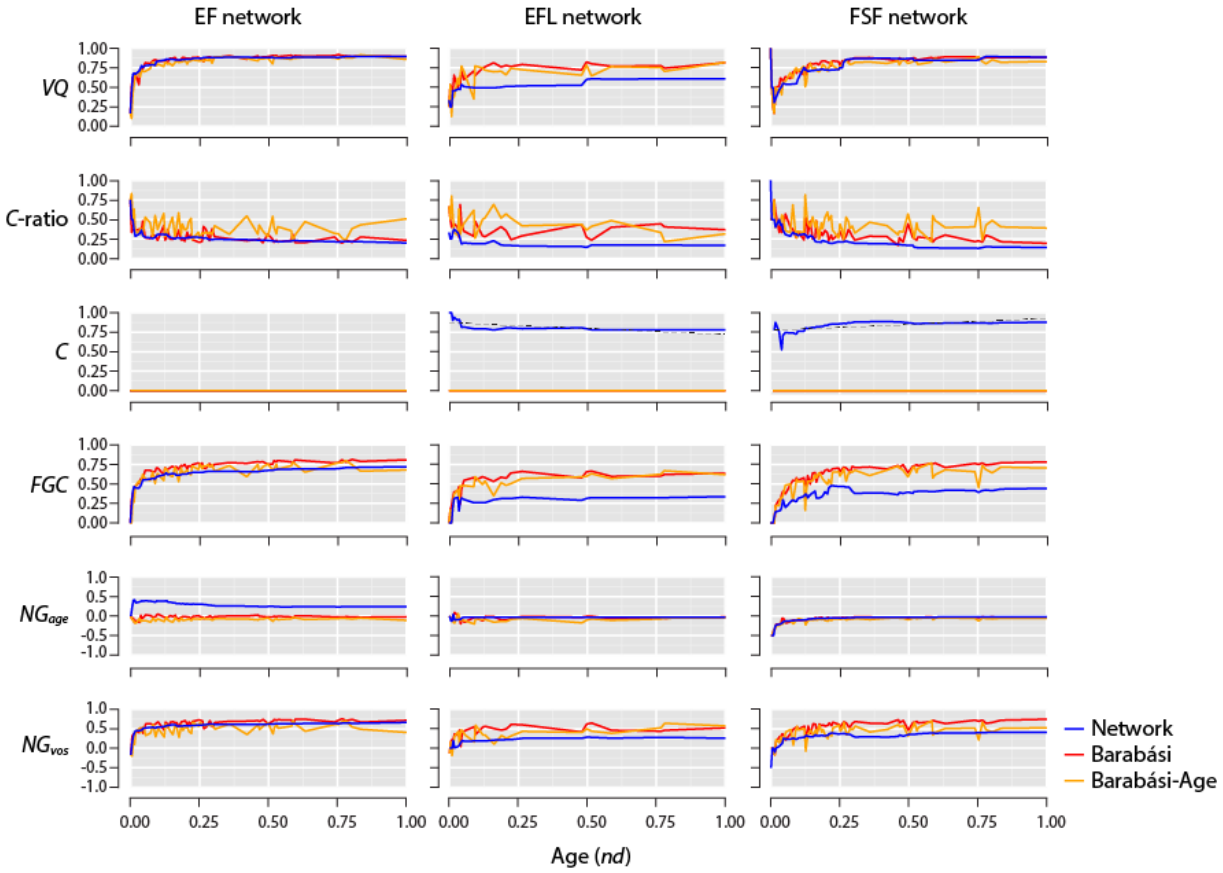

Figure 3.10. **Modularity indices of the EF bipartite network and its EFL and FSF network projections.** Six indicators of modularity were plotted for each network. The coefficients of linear regression lines (grey) over *C* for the FSF network were 0.0022 by network size (*N*) and 0.17 by age, and those for the EFL network were -0.006 by *N* and -0.15 by age. Determination coefficients (R^2) were 0.542, 0.369, 0.802 and 0.327, in that order. The linear models shown are by *nd*. Two modeled control sets, Barabási (red) and Barabási-Age (orange) were included as reference. The "Age" control enriches preferential attachment properties in older (smaller *nd*) nodes, inducing slightly different modularity curves. Single comprehensive modularity curves were computed for each network. Modularity calculations required cumulative and weighted connectivity input. Age (*nd*) is indicated on a relative 0-to-1 scale.

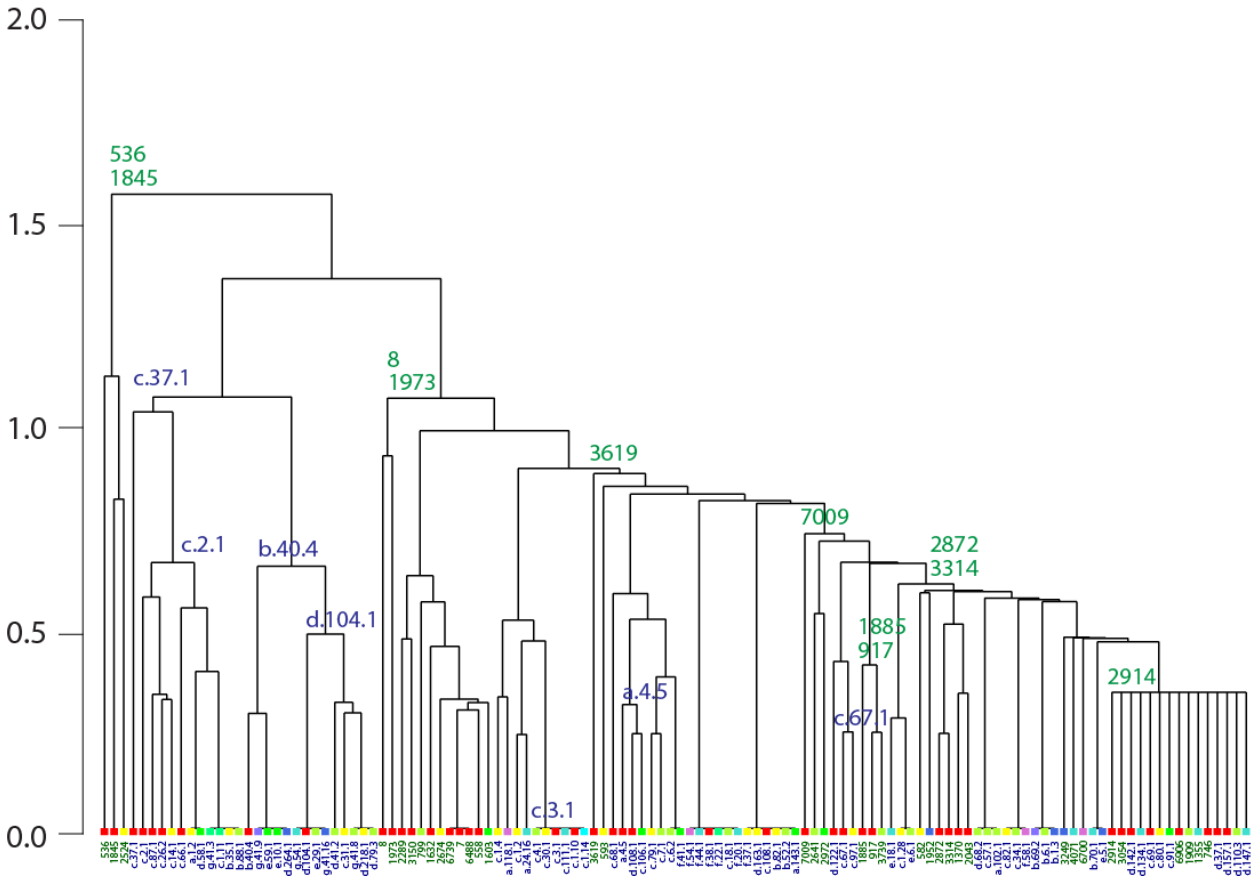

Figure 3.11. **Hierarchical organization of the fully-grown (extant) EF network (*nd*=1.000).** NG_{age} scaled modularity matrix (Newman and Girvan 2004) of the last EF network in the timeline was input to calculate Euclidean distance matrix (Borg and Groenen 2003), which was hierarchically clustered with the Ward's minimum variance method (Murtagh and Legendre 2014). Vertical axis represents dissimilarity between clusters of the dendrogram. Horizontal axis labels are color-coded to distinctively identify the rearrangements of the two classes of constructs. The nodes are age-sorted ascendingly within clusters, color-coded by node age and labelled using standard SCOP nomenclature (Murzin et al. 1995). The most ancient loop motifs and domain structures of the clusters comprising significant portions of the two major waves of functional innovation and the hidden EF switch of power law and modularity exchange are displayed on the roots of clades.

REFERENCES

Aharonovsky E, Trifonov EN. 2005. Protein sequence modules. *J Biomol Struct Dyn* 23(3): 237-242.

Albert R, Barabási AL. 2002. Statistical mechanics of complex networks. *Rev Modern Phys* 74(1):47.

Aziz MF, Chan P, Osorio JS, Minhas BF, Parekatt V, Caetano-Anollés G. 2012. Stress induces biphasic-rewiring and modularization patterns in metabolomics networks of *Escherichia coli*. *IEEE Intl. Conf. Bioinf. Biomed.* p593-597. doi: 10.1109/BIBM.2012.6392626

Barabási AL, Albert R. 1999. Emergence of scaling in random networks. *Science* 286(5439):509-512.

Barrat A, Barthelemy M, Pastor-Satorras R, Vespignani A. 2004. The architecture of complex weighted networks. *Proc Natl Acad Sci USA* 101(11):3747-3752.

Batagelj V, Mrvar A. 1998. Pajek-program for large network analysis. *Connections* 21(2): 47-57. From http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

Berezovsky IN, Grosberg AY, Trifonov EN. 2000. Closed loops of nearly standard size: common basic element of protein structure. *FEBS Lett* 466: 283–286.

Berezovsky IN, Trifonov EN. 2001. Van der Waals locks: loop-n-lock structure of globular proteins. *J Mol Biol* 307: 1419–1426.

Borg I, Groenen P. 2003. Modern multidimensional scaling: theory and applications. *J Educ Measurement* 40(3):277-280.

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.

Caetano-Anollés G, Kim HS, Mittenthal JE. 2007. The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proc Natl Acad Sci USA* 104: 9358–63.

Caetano-Anolles G, Wang M, Caetano-Anolles D, Mittenthal JE. 2009a. The origin, evolution and structure of the protein world. *Biochem J* 417:621–637.

Caetano-Anollés G, Yafremava LS, Gee H, Caetano-Anollés D, Kim HS, Mittenthal JE. 2009b. The origin and evolution of modern metabolism. *Intl J Biochem Cell Biol* 41:285–297.

Caetano-Anollés G, Kim KM, Caetano-Anollés D. 2012. The phylogenomic roots of modern biochemistry: Origins of proteins, cofactors and protein biosynthesis. *J Mol Evol* 74: 1-34.

Caetano-Anollés G, 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.

Chung FRK, Erdös P, Spencer J. 1983. On the decomposition of graphs into complete bipartite subgraphs. In: Erdös P, Alpar L, Halasz G, Saeközy A, editors. *Studies in pure mathematics, To the memory of Paul Turán*. Budapest: Bikhäuser Verlag. p95-101

Clauset A, Newman ME, Moore C. 2004. Finding community structure in very large networks. *Phys Rev E* 70(6):066111.

Clauset A, Shalizi CR, Newman ME. 2009. Power-law distributions in empirical data. *SIAM Rev* 51(4): 661-703.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. *Intl J 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. New York: Marcel Dekker Inc.

Diestel R. 2010. Graph theory. Graduate Texts in Mathematics, Vol. 173, Fourth Edition, Heidelberg: Springer-Verlag.

Goncearenco A, Berezovsky IN. 2010. Prototypes of elementary functional loops unravel evolutionary connections between protein functions. *Bioinformatics* 26: i497-i503.

Goncearenco A, Berezovsky IN. 2012. Exploring the evolution of protein function in Archaea. *BMC Evol Biol* 12: 75.

Edwards H, Abeln S, Deane CM. 2013. Exploring fold space preferences of new-born and ancient protein superfamilies. *PLoS Comput Biol* 9:1003325.

Ihaka R, Gentleman R. 1996. R: A language for data analysis and graphics. *J Comput Graph Stat* 5(3):299-314.

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. 2000. The large-scale organization of metabolic networks. *Nature* 407(6804):651-654.

Kim KM, Caetano-Anollés G. 2010. Emergence and evolution of modern molecular functions

inferred from phylogenomic analysis of ontological data. *Mol Biol Evol* 27: 1710-1733.

Kim HS, Mittenthal JE, Caetano-Anollees G. 2013. Widespread recruitment of ancient domain structures in modern enzymes during metabolic evolution. *J Integr Bioinform* 10(1):214.

MacDougall MH. 1987. Simulating computer systems: Techniques and tools. Cambridge: MIT Press.

Mittenthal J, Caetano-Anollés D, Caetano-Anollees G. 2012. Biphasic patterns of diversification and the emergence of modules. *Front Genetics* 3:147.

Murtagh F, Legendre P. 2014. Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *J Classification* 31(3):274-295.

Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology 247*(4):536-540.

Nath N, Mitchell JB, Caetano-Anollés G. 2014. The natural history of biocatalytic mechanisms. *PLoS Comput Biol* 10(5):e1003642.

Newman ME. 2005. Power laws, Pareto distributions and Zipf's law. *Contemp Phys* 46(5): 323-351.

Newman ME, Girvan M. 2004. Finding and evaluating community structure in networks. *Phys Rev E* 69(2):026113.

Newman ME, Strogatz SH, Watts DJ. 2001. Random graphs with arbitrary degree distributions and their applications. *Phys Rev E* 64(2):026118.

Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov Jr E, Kyrpides N, Selkov E. 2000. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res* 28(1):123-125.

Pidd M. 2004. Computer simulation in management science. John Wiley & Sons, New York.

Pons P, Latapy M. 2006. Computing communities in large networks using random walks. *J Graph Algorithms Appl* 10(2):191-218.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. 2002. Hierarchical organization of modularity in metabolic networks. *Science* 297(5586):1551-1555.

Sobolevsky Y, Guimaraes RC, Trifonov EN. 2013. Towards functional repertoire of the earliest proteins. *J Biomol Struct Dyn* 31(11):1293-1300.

Strogatz SH. 2001. Exploring complex networks. *Nature* 410: 268-276.

Teichmann M, Dumay-Odelot H, Fribourg S. 2012. Structural and functional aspects of the winged-helix domains at the core of transcription initiation complexes. *Transcription* 3:1,

Trifonov EN, Frenkel ZM. 2009. Evolution of protein modularity. *Curr Op Struct Biol* 18:335-340.

Van Eck NJ, Waltman L. 2007. 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*. Heidelberg: Springer Verlag. p299-306.

Wagner A, Fell DA. 2001. The small world inside large metabolic networks. *Proc Roy Soc London Series B: Biol Sci* 268(1478):1803-1810.

Waltman L, Van Eck NJ, Noyons EC. 2010. A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, *4*(4), 629-635.

Wang M, Jiang YY, Kim KM, Qu G, Ji HF, Mittenthal JE, Zhang HY, 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. *Mol Biol Evol* 28:567-582.

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.

Wasserman S, Faust K. 1994. Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.

APPENDIX A: VIDEOS

Video 1. Simulation of pairwise NG_{age} modularity of the EF network through progression of heat maps over the timeline, May 26, 2015. The video may be found in a supplemental file named **Video 1 EFL_FSFHeatmapNG_age.mp4**.

Video 2. Simulation of hierarchical organization of the EF network based on NG_{age} modularity through progression of dendrograms over the timeline, May 26, 2015. The video may be found in a supplemental file named **Video 2 EFL_FSFDendrogramNG_age.mp4**.

APPENDIX B: LIST OF ACRONYMS

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
EF	Elementary Functionome
EFL	Elementary Functional Loop
FGC	Fast Greedy Community detection algorithm
FSF	Fold SuperFamily
Gy	Geological years (billions of years)
KS	Kolmogorov-Smirnov statistical test of power law fit
LM	Linear Model
Log L	log Likelihood
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

SCOP	Structural Classification of Proteins
VOS	Visualization of Similarity clustering algorithm
VQ	VOS Quality index
WTC	Walk Trap Community detection algorithm