A PHYLOGENOMIC HISTORY OF PROTEIN FUNCTION AND DYNAMICS

Allocation: Illinois/200 Knh
PI: Gustavo Caetano–Anollés[1]
Co-PI: Frauke Gräter[2]
Collaborator: Fizza Mughal[1]

[1]University of Illinois at Urbana–Champaign
[2]Heidelberg Institute for Theoretical Studies

EXECUTIVE SUMMARY

Studying the evolution of protein function is important for synthetic biology and translational medicine. The ability of proteins to undergo motions that perform a certain function depends on their structural flexibility. Flexibility, an evolutionarily conserved feature of protein structure, is a suitable proxy to measure protein dynamics and the underlying evolutionary drivers that sustain molecular function. In this work, the research team investigated the presence of evolutionary constraints on protein motions using molecular dynamics (MD) simulations, generating dynamics networks that capture topological features of protein structures, and constructing a three-dimensional network morphospace to trace the evolutionary emergence of protein function.

RESEARCH CHALLENGE

Proteins perform a multitude of functions that sustain life on our planet. Understanding their evolution can impact agriculture, bioengineering, and biomedicine. Protein loops play an important role in protein function and dynamics by virtue of their structural flexibility [1]. Conservation of protein dynamics and flexibility [2] suggest the existence of signature motions corresponding to each protein function. Therefore, probing properties of loops at higher and lower levels of cellular organization may provide evolutionary clues of how protein function shapes protein dynamics. Previous work from the research group has demonstrated the utility of networks to model evolutionary interaction at the organizational level of protein domains and loops [3].

The team is extending this methodology at the protein loop and residue level to study biophysical properties of associated functions, combining the seemingly disparate fields of physics and evolution, and leveraging nanosecond dynamics to dissect billions of years of phylogenomic history.

METHODS & CODES

The research team extended its previous data set of 116 loops from protein domains belonging to metaconsensus enzymes [4] by including 58 additional structures. This augmented the group's structural data set with previously underrepresented functional categories. The all-atom MD simulations were performed using an isobaric–isothermal ensemble (NPT) in TIP3P (transferable intermolecular potential with three points) water. We applied harmonic restraints of 2.1 kcal/mol Å^2 to the bracing secondary structure of the peptide. A sodium and chloride ionic concentration of 100 mMol was used to mimic near-physiological conditions. Depending on the number of atoms in the system, the researchers performed 50 to 70 ns (nanosecond) production runs with 1 ns of minimization using NAMD and the CHARMM36 force field. They generated networks based on the dynamic cross-correlation matrices computed from the simulations, from which they calculated network metrics for cohesion and centralities using the R packages bio3D and igraph, respectively [5,6]. The research team performed community structure detection on these networks that in turn generated trees for which they computed imbalance metrics using the R package phyloTop [7]. Other than network metrics, variables capturing the biophysical properties of the structure and corresponding movements were directly measured using principal component analyses, radius of gyration, and root mean square deviation. The protein structures were annotated with evolutionary age (nd) derived from phylogenomic timelines developed in the research team's lab [8].

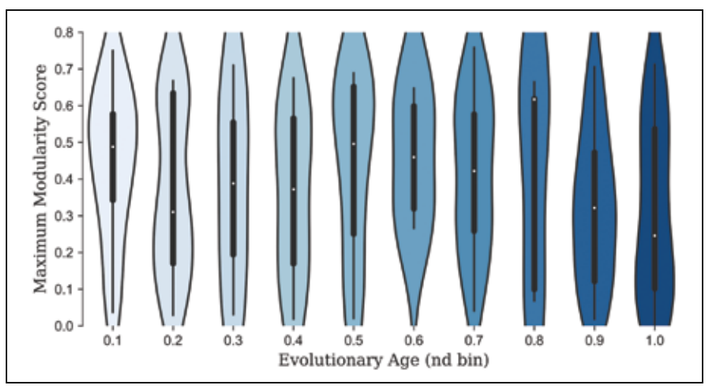

Figure 1: The maximum modularity scores (y-axis) for 170 of the 174 dynamic networks across the evolutionary timeline of protein domains (x-axis) spread across 3.8 billion years of evolution. Violin plots describe measures of central tendency with box-and-whiskers depictions of medians, quartiles, and data spread, all of them embedded within kernel density plots of the data. Modularity is high at the beginning of the timeline, and decreases with time with episodic up-down fluctuations.

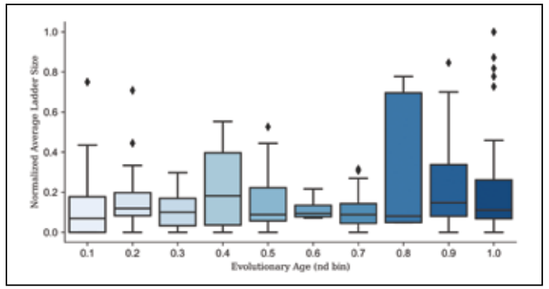

Figure 2: The average ladder size (y-axis) for trees based on community network structures for 170 of the 174 MD simulations distributed across the evolutionary timeline (x-axis). Box-and-whiskers plots describe how average ladder size stayed mostly consistent and closer to the mean for each age (nd) bin. However, a biphasic pattern was evident with average ladder sizes positively skewed late in evolution (especially at nd bin = 0.8). Skews indicate higher values than the median for the nd bin. Rhomboid symbols represent outliers.

In addition to MD simulations, the researchers leveraged comparative genomics to assess whether protein loop structures hold significant evolutionary signal. Nearly 2,100 proteomes belonging to Archaea, Bacteria, and Eukarya along with 6,044 viral proteomes were downloaded from the RefSeq database [9]. The team included proteomes from representative and reference categories with chromosome or complete genome assembly in the study, as well as all viral proteomes listed in the National Center for Biotechnology Information viral genomes project [10]. The final set of proteomes was scanned against HMM profiles of structural domains using HMMER [11]. This genomic survey of protein domains was used to generate feature matrices to construct maximum parsimony trees of domains, following the protocol established by Kim *et al.* [8]. The resulting tree of domains with proteomes as characters are being compared against tree of domains with protein loop architectures as characters. The congruency of both the trees will help in establishing the presence (or absence) or phylogenomic signal embedded in loop architectures carried by the protein domains.

RESULTS & IMPACT

Topological metrics of dynamic networks generated from the MD simulations provide interesting insight into molecular organization at a secondary-structure level. Network modularity decreased with time along the evolutionary timeline (Fig. 1), with episodic fluctuations embodying a biphasic model of module innovation and growth [12]. In the beginning, parts of a system have weak linkages. These weak linkages instigate diversification of parts through mutation, recruitment, and reassortment. Following diversification, competitive optimization occurs, leading to a decrease in diversity and, subsequently, hierarchical organization of modules. The modules undergo diversification again to give rise to novel modules at yet another level of organization [12].

Phylogenetic trees are routinely used in epidemiology and evolutionary biology to study transmission patterns [13]. Quantifying the asymmetric nature of trees may help explain growth and innovation of biological features and the hierarchical structure of networks. In this study, the researchers constructed trees from communities in dynamic networks to investigate transmission of modular expansion in protein dynamics.

They measured tree asymmetry using a number of statistics, of which one is the average ladder size. A ladder is a series of nodes linked to a common ancestor, with each node having exactly one tip child (leaf) [14]. The average of all ladders in a tree was measured and normalized on a scale of 0 (balanced) to 1 (highly imbalanced). Average ladder sizes tend to be low and not as widespread for each age (nd) bin (Fig. 2). However, the average ladder sizes follow a biphasic pattern that tends to be positively skewed (above the median value) during the second phase (especially at age bin = 0.8), indicating that relatively recent dynamic networks have hierarchies that are more imbalanced than dynamic networks of protein loops from older domains.

WHY BLUE WATERS

The computational heavy lifting of Blue Waters facilitated the completion of time-intensive research endeavors, including the scanning of the proteomes of approximately 2,000 organisms and thousands of viruses with sophisticated hidden Markov models of protein domain recognition. Other computational experiments involved MD simulations of 300 all-atom explicit water peptide systems. Access to Blue Waters helped the research team to complete these studies in a reasonable time period. Blue Waters support staff were knowledgeable in supercomputing matters and comprised domain experts. They were extremely helpful in answering both technical and field-specific queries.

PUBLICATIONS & DATA SETS

G. Caetano-Anollés *et al.*, "Emergence of hierarchical modularity in evolving networks uncovered by phylogenomic analysis," *Evol. Bioinform.*, vol. 15, no. 1, p. 1176934319872980, Sept. 2019. doi: 10.1177/1176934319872980