GA

ML

HOW FUNCTION SHAPES DYNAMICS IN PROTEIN EVOLUTION

Allocation: Illinois/350 Knh
PI: Gustavo Caetano-Anolles[1]
Co-PIs: Frauke Gräter[2]
Collaborators: Fizza Mughal[1]

[1]University of Illinois at Urbana-Champaign
[2]Heidelberg Institute for Theoretical Studies (HITS)

EXECUTIVE SUMMARY

Protein loops are found to be chiefly responsible for the wide functional diversity of proteins. This stems mostly from the property of protein flexibility, which is evolutionarily conserved. This suggests specific molecular motions may have been selected for individual functions. By harnessing the power of Blue Waters, we aim to decipher patterns and processes underlying the origin, evolution, and structure of the molecular dynamics of proteins. In our current allocation, we have completed 116 molecular dynamics simulations of loop regions of protein structural domains found in metaconsensus enzymes. In addition, we completed aminoacyl-tRNA synthetase (aaRS) simulations that were pending from a previous allocation. The collected data were subjected to a preliminary analysis of molecular trajectories. Variables were computed that described the dynamic properties in these trajectories permitted to construct a dynamics space, a "dynamosome," that we intend to map onto a "structure–evolution" protein space.

RESEARCH CHALLENGE

The biophysical properties of protein loops may hold answers to the discovery, prediction, and annotation of protein functions [1]. Some of these biophysical properties may be governed by yet-to-be discovered evolutionary drivers that could significantly impact synthetic biology and translational medicine [2]. Incorporation of biophysics in protein structure–function studies is becoming increasingly common. However, biophysics is rarely used in evolutionary studies [3]. The main objective of our studies is to bridge disparate disciplines of biology and physics with Molecular Dynamics (MD) simulations performed at nanosecond (ns) timescales to capture evolutionary dynamics on a scale of billions of years. Here, we explore biophysical variables of the MD simulations by studying community structures of protein loop residues that describe the molecular trajectories of the loop regions. Our goal is to dissect evolutionary relationships in these data using evolutionary timelines reconstructed from robust phylogenomic methods [4].

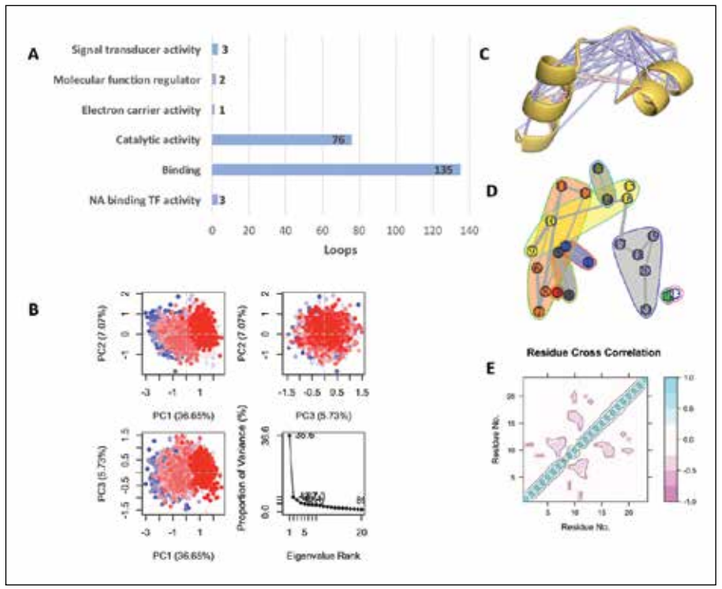

Figure 1: A: Distribution of functional annotations across 86 of 87 loops that exhibited a community dynamics structure. A total of 72 out of the 87 loops examined possessed (single or multiple) Gene Ontology (GO) functional annotations. B: Principal Component Analysis of Protein loop 1B7Y_B_408 associated with the a.6.1.1 SCOP domain. C: Protein loop 1B7Y_B_408; residues connected to each other due to positive (red) and negative (blue) correlations of motions during MD trajectory. D: All-residue network of 1B7Y_B_408 with community structures highlighted as groups. E: Dynamical Cross Correlational Map of the protein residues of 1B7Y_B_408. Featured in *Blue Waters Annual Report* (Urbana, Illinois, 2016).

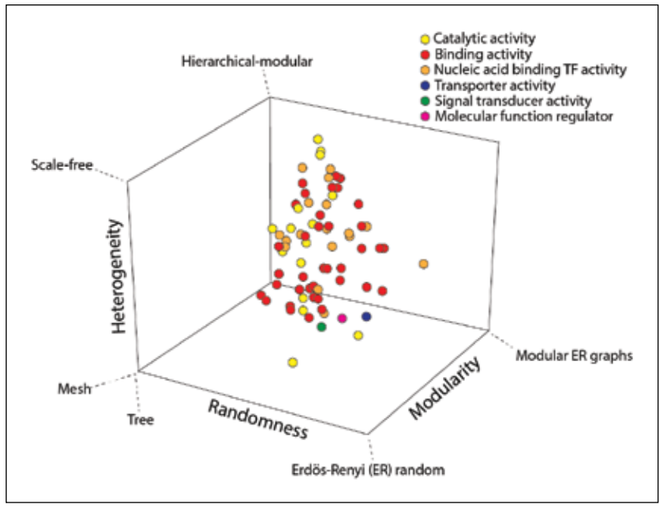

Figure 2: The community dynamics network morphospace of aaRS protein. A Pareto front defines a Zooko-like triangle of trade-off solutions among economy, robustness and flexibility for protein dynamics.

METHODS & CODES

We completed MD simulations that were left pending in a previous allocation, in which we analyzed 87 protein loops from aaRS structural domains on a timescale of 50–70 ns. In addition, we simulated 116 protein loops belonging to single-domain metaconsensus enzymes. The protein loops of aaRS domains were mostly associated with the Gene Ontology (GO) level-1 molecular function of "binding" followed by that of "catalytic activity" (Fig. 1.A). We constructed a dynamics space, a modified version of the dynamosome [5], by calculating the eigenvalues of the top five principal components from principal component analysis (Fig. 1.B) and centrality metrics from a network (Fig. 1.D) based on the dynamic cross-correlation matrix of the motions of protein residues (Fig. 1.E). In order to assess the presence or absence of a specific network topology, we also calculated maximum modularity scores, alpha values to test power law behavior, and Bartel's test statistic for measuring the extent of modularity, scale-freeness, and randomness of the network (Fig. 2). We are currently in the process of performing unsupervised clustering of the trajectories using the dynamosome variables. We also plan to use methods that classify community structure patterns (Fig. 1.D) exhibited by loops and their correlation to specific function. Our goal is to reconstruct a "structure–evolution" space that would complement our dynamosome.

RESULTS & IMPACT

The aim of our investigation is to detect the presence or absence of patterns of motion in molecules. We focus on an analysis of dynamic network topologies defining a three-dimensional morphospace delimited by the conceptual axes of modularity, scale-freeness, and randomness [6]. Modularity, a feature persistently observed in biological networks [7], embodies flexibility and diversity of the molecular components that make up the whole. Scale-freeness is an indicator of heterogeneity in patterns of connectivity of the network. It is a measure of "economy" (i.e., how easy it is to traverse the network structure). Randomness entails uniform connectivity of nodes throughout the network, a property that confers network fault tolerance. Fig. 2 shows network topology tendencies for 72 of the 87 protein loops that have been annotated with GO functions. This plot encapsulates trade-offs among flexibility, economy, and robustness that result in a "noisy" 2-polytope Pareto front. Interestingly, Zooko's triangle, a concept used in the design of internet domain-name systems [8], can be thought to illustrate the design space of community networks obtained from protein dynamics. Protein loops possessing various functions tend to cluster at the center of this "triangle" (Fig. 2). They also prefer enhancing modularity in their quest to seek temporal persistence.

WHY BLUE WATERS

The petascale competencies of Blue Waters have been of great value to our project. With the help of such a state-of-the-art system, we have been able to achieve our goal of simulating a significantly large number of proteins (each at a timescale of ~70 ns) in a short time period. NAMD scales well on the Blue Waters architecture, especially when combined with GPU (graphics processing unit) nodes. This provides a significant boost in acceleration [9]. Apart from system specifications that are well-suited to our project, the domain experts/scientists in the Blue Waters support team have helped us smooth out any technical issues that have arisen during the current and previous allocations.