DI

BI

UNTANGLING THE ORIGINS OF PROTEIN FLEXIBILITY AND BIOLOGICAL FUNCTIONS

Allocation: Illinois/200 Knh
PI: Gustavo Caetano-Anollés[1]
Co-PIs: Frauke Graeter[2], Edmond Lau[3]
Collaborators: Fizza Mughal[1]

[1]University of Illinois at Urbana-Champaign
[2]Heidelberg Institute for Theoretical Studies
[3]Lawrence Livermore National Laboratory

EXECUTIVE SUMMARY

Protein loops are flexible elements of a macromolecular structure that are responsible for the diverse repertoire of biological functions of a proteome, or the complete assortment of proteins expressed by an organism. The flexibility of protein loops is an evolutionarily conserved molecular property that is critical for the dynamics of proteins. Employing dynamic networks to summarize the atomic trajectories of loops of metabolic enzymes obtained from 300-nanosecond-long molecular dynamics (MD) simulations, we study how network structure changes with molecular functions and protein history. We first scanned ~2,100 representative proteomes with Hidden Markov Model profiles of domain structure. We then used this census to generate evolutionary timelines of domains with which to annotate the networks. Finally, we analyzed the structure of the networks and their associated functions. We uncovered both robustness and strong biases in the dynamic trajectories of our simulations over billions of years of deep evolutionary time.

RESEARCH CHALLENGE

Protein loops are promising candidates for uncovering the evolutionary relationship between protein function and dynamics, an important topic that is still poorly understood. Loops are irregular secondary structures that account for the bulk of the molecular flexibility of the three-dimensional structure of proteins. They can be considered critical parts of the dynamic personality of proteins [1]. Protein dynamics is intricately related to protein structure. It has been hypothesized that dynamics "preexists" and shapes the evolution of proteins as they adapt to carry out specific sets of motions [2]. Additionally, flexibility has been found to be conserved in protein evolution [3]. Our study tries to make the link between dynamics and evolution by: (1) exploring whether form indeed follows function and (2) uncovering the evolutionary drivers responsible for shaping the dynamics of proteins. In our previous studies, we looked at this problem from a biophysical standpoint. Here, we uniquely coupled the nanosecond dynamics of molecules to a historical study of the function–dynamics relationship of proteins spanning billions of years of evolution.

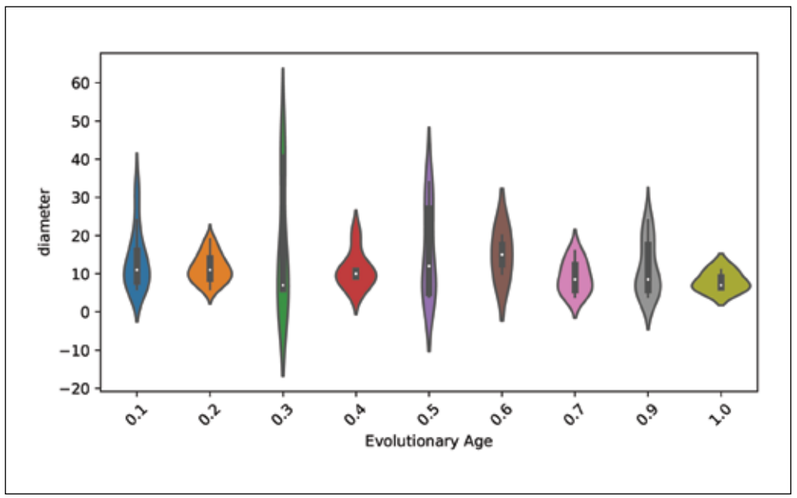

Figure 1: The diameter (y-axis) of dynamic networks across the evolutionary timeline (x-axis) of structural domains spanning 3.8 billion years of evolution. Evolutionary age ranges from the origin of proteins (0) to the present (1). The mean of distributions suggests network size is maintained during protein evolution.

METHODS & CODES

Molecular dynamics simulations. We studied 116 candidate loops that belong to the protein domains present in meta-consensus enzymes [4]. Our data set represents the entire set of seven broad categories of functional classification of structural domains. The MD simulations employed an isobaric–isothermal ensemble, TIP3P water model, harmonic restraints of 2.1 kcal/mol Å^2 applied to the bracing secondary structure of the protein, and a 100-mmol concentration of sodium and chloride ions. We performed the 50–70 nanosecond (ns) production runs preceded by 1 ns minimization runs using NAMD with the CHARMM36 force field.

Comparative genomic methods. We used the RefSeq database [5] to shortlist proteomes belonging to archaea, bacteria, eukarya, and viruses on the basis of the following two criteria: (1) organisms were classified as part of either the "representative" or "reference" RefSeq categories; and (2) the genome assembly was referred to as either "complete" or "chromosome."

Unclassified and misclassified organisms were removed from the resulting data set. In addition, we excluded organisms that have an exclusively "obligate" lifestyle (such as endosymbionts or phytoplasma). Such organisms tend to have small genomes and thus possess a limited set of protein domains, which distort phylogenetic relationships [6]. We applied this criterion to organisms from the three superkingdoms but not to viruses in order to have representation of viral domains in our data set. In cases when multiple subspecies or strains were present in the data set, we chose one organism only.

We scanned the resulting ~2,100 proteomes with protein domain HMM profiles using HMMER [7]. The results from HMMER scans are necessary to place loop-domain distribution across the superkingdoms in the historical context by mapping the loop classifications [8] against phylogenomic timelines developed in our lab [9]. We also studied the features of these protein domains in conjunction with loop dynamics. For this purpose, we used a nonredundant set of ~14,000 representative proteins [10]. The domain features were calculated in the form of protein blocks (conformational prototypes) [11], while the loop features were determined by the extent of its dynamicity, i.e., whether the loops are "static," "slow," or "fast" [12].

Data mining. To analyze the MD simulations of protein loops in single-domain meta-consensus enzymes, we generated dynamic networks of positive and negative correlations of motions based on these simulations, which we term "dynamic networks," and calculated important network metrics that measure cohesion and centralities. We then constructed a dynamics morphospace based on network metrics as well as principal component analyses and structural properties such as radius of gyration and Root Mean Square Deviation values.

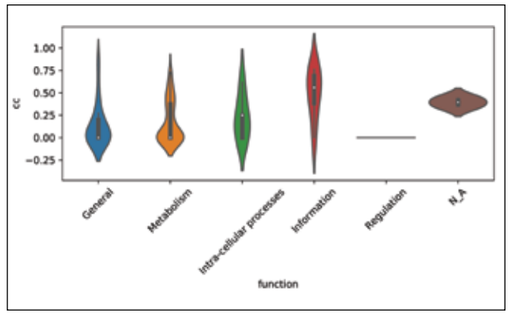

Figure 2: Distribution of clustering coefficients (cc, y-axis) of dynamic networks indexed according to functional categories of superfamilies (x-axis). The mean of the clustering coefficient belonging to the "information" and "not annotated" categories is significantly higher than that of the other categories.

RESULTS & IMPACT

A global study of the dynamic networks observed in our MD simulations suggests remarkable leads for more detailed exploration. The average values of the diameter and length of the dynamic networks along the evolutionary timeline of structural domains remained constant throughout the entire 3.8 billion years of evolution (Fig. 1). This finding strongly suggests that dynamics is an entrenched physical property of proteins. When studying the clustering coefficient of dynamic networks that were annotated according to Vogel's functional classification of structural domains [13], the mean of the clustering coefficient distribution of structural domains belonging to the "information" and "not annotated" general functional categories was significantly higher than those of the rest of the categories (Fig. 2). The "information" category encompasses domains that play a role in the upkeep and storage of the genetic material. Additionally, they are involved in information flow (transcription and translation) as well as replication and repair processes of the nucleic acids. Higher clustering coefficients are believed to correspond to higher levels of modularity in network structure. That, in turn, suggests informational processes in biology are compartmentalizing molecular dynamics into distinct but cohesive behaviors. Our analyses suggest that robustness in the structure of dynamic networks is tempered by the emergence of modules of dynamic behavior in specific functions of the molecular systems.

WHY BLUE WATERS

Blue Waters enabled the completion of MD simulations of loop behavior in 300 molecules and the scanning of a vast number of proteomes with advanced HMMs of structural recognition. Without access to Blue Waters, this computationally intensive study would not have been possible to achieve in a reasonable timeframe. We commend the Blue Waters support staff. They have been extremely helpful with prompt resolution of computational and other logistical issues during the execution of this project.