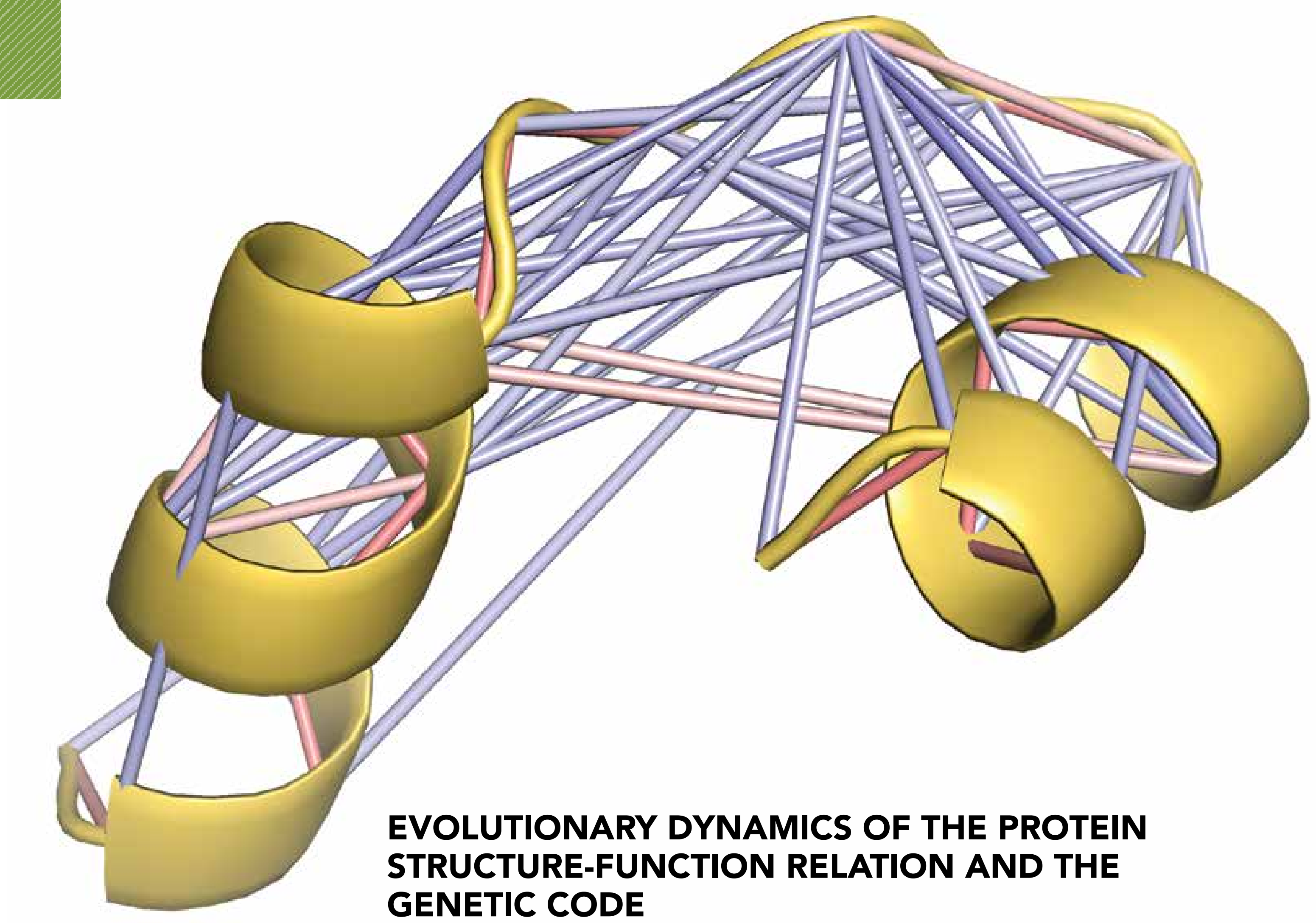

FIGURE 1: Protein loop 1B7Y_B_408 associated with the a.6.1.1 SCOP domain with residues connected to each other based on positive (red) and negative (blue) correlations of motions during the MD simulation.

EVOLUTIONARY DYNAMICS OF THE PROTEIN STRUCTURE-FUNCTION RELATION AND THE GENETIC CODE

Allocation: Illinois/350 Knh
PI: Gustavo Caetano-Anolles[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

The molecular functions of a protein are determined by the dynamics of its flexible structural components, also known as protein loops. These loop regions exhibit distinct sets of molecular motions, which associate with specific functions. Here we study the biophysics of loop motions in protein structural domains with ages spanning the entire timeline of protein evolution. We are currently working on the molecular dynamic simulations of an initial set of 87 loops embedded within the aminoacyl-tRNA synthethase (aaRS) enzymes responsible for delimiting the specificity of the genetic code. We are also data mining the simulations with machine learning methods. Preliminary analyses using graph theoretical approaches reveal communities of synchronized movements preferentially associated with the secondary structure at the C-termini of the loop regions. These analyses have extended to the sampling of metaconsensus metabolic enzymes and could unravel hidden evolutionary links between the dynamics and functions of proteins.

INTRODUCTION

Protein loops are unstructured regions that play an important role in defining the function and structural stability of proteins [1]. Loops are not completely disordered. They endow an individual protein with characteristic motions that are vital to the protein's function. Protein dynamics, along with backbone flexibility, have been found to be strongly conserved [2]. While the regular secondary structure is highly rigid, the irregular loop counterparts are major contributors to molecular flexibility. The correlation between protein dynamics and function is demonstrated by how certain non-homologous enzymes with similar molecular functions tend to display similar motions [2]. Also, the molecular functions of proteins have been attributed to a combination of functional activities enabled by an evolutionarily conserved set of "elementary functional loops" (EFLs) that associate to form active and regulatory sites in the molecules [3]. Here we investigate this structure-function paradigm using evolutionary dynamics. For this purpose, biophysical variables are selected as candidates for functional annotation of protein domain structures and their loop components [4]. The impact of both evolution and biophysics on the makeup of proteins may provide deeper insight into gains and losses of structural domain features of biological macromolecules [5]. The goal is to decipher the drivers of molecular evolution by reconstructing the evolutionary past. Our work has the potential to build evolutionary trajectories necessary for biomolecular engineering, synthetic biology, and translational medicine. For example, understanding molecular evolution can help uncover a rationale for "engineered" metabolic pathways or can be used to develop better viral vaccines by studying how mutations affect the structure and dynamics of capsid proteins [6].

METHODS & RESULTS

Previous work using Blue Waters revealed that molecular flexibility plays a major role in defining evolutionary constraints acting upon a protein [7]. A protein loop in our study refers to both the unstructured loop structure and the bracing secondary structures that define the return of the polypeptide backbone (Fig. 1). We have simulated 87 loops associated with the domains of aminoacyl-tRNA synthetase (aaRS) enzymes using NAMD 2.9 simulation with force field parameters specified by CHARMM36 files, each for a duration of 10 nanoseconds. These loops have been annotated using gene ontology (GO) molecular functions and classified using the ArchDB "Density Search" system.

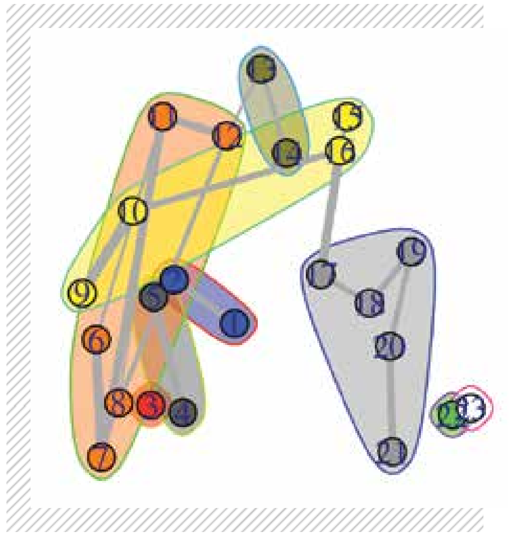

FIGURE 2: All-residue network of 1B7Y_B_408 with emphasis on community structures.

The molecular dynamics simulations were analyzed globally by computing parameters such as radius of gyration and root mean square deviation (RMSD). Also, local parameters were assessed by calculating the root mean square (RMS) fluctuations, exploring parameter variability with principal component analyses (PCA) and plotting community behavior in networks of motions. As expected, the residues in the unstructured parts of the loops possess the highest values of RMS fluctuation. The PCA is projected onto the candidate loop structure to ascertain the type of associated motion. The communities in our network analyses (Fig. 2) consist of residues that have synchronized movements during the simulation, based on residue cross-correlations (Fig. 2). Remarkably, the betweenness values plotted from the resulting networks showed high values for residues in the C-terminus secondary structure for the majority of structures that were examined. Future analyses will focus on uncovering patterns in the community structures, as well as motions, thereby dwelling deeper into the structure, function and dynamics of protein loops. These analyses will be followed by mapping patterns along an evolutionary timeline of loop-embedding structural domains generated with robust phylogenomic reconstruction methods [8].

WHY BLUE WATERS

Blue Waters has been instrumental in enabling the study of the structure-function protein interplay at the core of the origin of the genetic code. With the help of Blue Waters, we have successfully

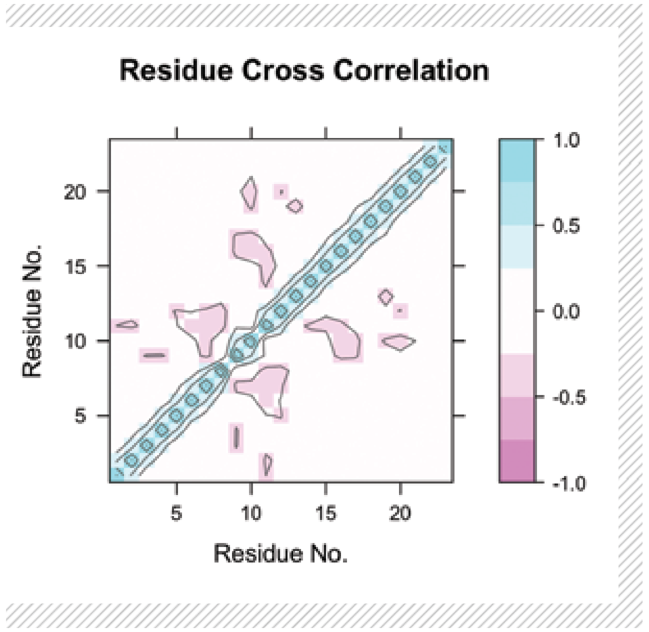

FIGURE 3: Dynamical Cross Correlational Map of the protein residue backbone of 1B7Y_B_408.

simulated a total of about 1,500 nanoseconds (1.5 milliseconds) worth of simulations involving 87 protein loop classifications associated with aaRS domains delimiting the specificities of the genetic code. Loop molecular dynamic simulations of these multi-domain proteins have yielded impressive results which have laid the **groundwork** for analyses involving single-domain metabolic metaconsensus enzymes selected using comparative bioinformatics techniques grounded in sequence, structure, and metabolic reactions.

NEXT GENERATION WORK

Our preliminary experiments indicate an intricate set of patterns among structure, function, and dynamics of proteins. These patterns have the potential to **uncover** evolutionary drivers that may shed light on a basic yet confounding phenomenon in nature: does structure dictate function or vice versa? We aim to expand our analysis to bigger datasets concerning proteins in signaling networks. Also, we are interested in performing machine learning analyses of dynamic simulation datasets to dissect distinct molecular dynamic patterns along an evolutionary timeline of protein domains.

PUBLICATIONS AND DATA SETS

Mughal, F., G. Caetano-Anollés and F. Gräter. Mining the evolutionary dynamics of protein loop structure and its role in biological functions. *Blue Waters Annual Report* (Urbana, Illinois, 2015), pp.130-131.

Caetano-Anollés, G., F. Gräter, C. Debès, D. Mercadante, and F. Mughal. The dynamics of protein disorder and its evolution: Understanding single molecule FRET experiments of disordered proteins. *Blue Waters Annual Report* (Urbana, Illinois, 2014), pp. 100-101.

IMPROVING THE ACCURACY OF DRUG PERMEABILITY CALCULATIONS

Allocation: Illinois/25.0 Knh
PI: Christopher Chipot[1]
Co-PI: Jeffrey Comer[2]

[1]University Illinois at Urbana-Champaign
[2]Kansas State University

EXECUTIVE SUMMARY

The inhomogeneous solubility-diffusion model has provided a convenient framework for understanding membrane permeation by drug molecules. This model shows the relationship between the resistance to permeation in the direction normal to the membrane to the position-dependent diffusivity of the drug and the one-dimensional free-energy profile underlying its translocation from the bulk aqueous phase to the interior of the lipid environment. For the **first time**, we provide a model for membrane permeation of a drug that, in stark contrast with the solubility-diffusion model, does not assume a lack of long-range correlations in time and space. Our model allows for better understanding of permeation dynamics for molecules exhibiting subdiffusive behavior on the characteristic timescales of their permeation. Our simulations suggest that this subdiffusive behavior is a result of permeation being governed by the spontaneous formation of voids within the membrane, which leads to intermittent large displacements of a permeant that is otherwise nearly immobile.

INTRODUCTION

In the search of novel therapeutic agents, many chemical compounds able to bind a given target with very high affinity are eventually discarded on account of their cytotoxicity, propensity to associate with potassium channel human Ether-à-go-go-Related Gene (hERG), or poor bioavailability. Predicting these properties at an early stage of drug discovery, upstream from costly organic syntheses and clinical trials, is desirable. One possible avenue to address high drug-attrition rates [1] consists in quantifying the ability of the substrate to traverse lipid membranes spontaneously, for instance, in the gastrointestinal tract, and reach the targeted protein in an adequate amount. A consistent theoretical model of the lipid membrane permeation process is essential for linking the physicochemical properties of drug candidates to their adsorption and distribution.

METHODS & RESULTS

The goal of this research is to understand a question central to drug discovery, namely how a drug spontaneously translocates across the biological membrane to reach its designated target. A model that has pervaded the field over the past twenty years is the so-called solubility–diffusion model of passive membrane permeation of small molecules [2]. In this model, the diffusion of the permeant is ordinarily assumed to obey the conventional Smoluchowski diffusion equation, which describes classical diffusion of particles on an inhomogeneous free-energy and diffusivity landscape. However, this equation cannot accommodate subdiffusive behavior [3], which has long been recognized in other aspects of lipid bilayer dynamics, including lateral diffusion of individual lipids. Using large-scale molecular dynamics simulations of permeation events in a fully hydrated lipid bilayer performed on Blue Waters, we show that subdiffusive behavior is present in the transverse diffusion of a series of alcohols through a pure membrane, remaining relevant on timescales approaching the typical permeation time. We find that a model based on a fractional-order differential equation appropriately describes the motion of the permeant on timescales ranging from 1 picosecond to 1 nanosecond, which cannot be replicated by a single conventional Smoluchowski model. Multiple approaches indicate that the mean squared displacement within the bilayer, in the absence of a net force, depends on time as a power law, namely $t^{0.7}$, in contrast with the conventional model where this dependence is strictly linear. Our molecular dynamics simulations bring to light an unexpected phenomenon, linking subdiffusion to the formation of transient voids that spontaneously appear within the hydrophobic region of the bilayer and allow rare, but large displacements of the permeant, which is otherwise virtually immobile. The results of this investigation, which reweaves the fabric of the physical principles underlying membrane permeation by drug molecules, have been reported in a research article recently submitted for publication [4].

WHY BLUE WATERS

Blue Waters was essential to perform a very large series of independent molecular-dynamics simulations of a membrane assembly in a time-bound fashion.