***MB&B***

Yale University

***260/266 Whitney Avenue***

***PO Box 208114***

***New Haven, CT 06520-8114***

***Telephone:***

***203 432 6105
360 838 7861 (fax)***

***mark@gersteinlab.org***

***www.gersteinlab.org***

Sept. 17 2015

Dear Simons Foundation Program Staff and Selection Committee,

I write to express my great interest in applying to the Simons Foundation Targeted Grants in the Mathematical Modeling of Living Systems program. On the subsequent pages, I have outlined my research plans for a project that will [please fill in].

To date, my research group has worked extensively to analyze large-scale protein conformational changes. This work has provided novel insights into [please fill in as needed] and has resulted in the development and dissemination of several widely used online software tools for analyzing and visualizing protein structures and motions (molmovdb.org). Given the ubiquity and importance of allosteric regulation, we now aim to leverage our expertise to develop new methods that will enable the field to gain new insights into long-standing questions such as, [please fill in as needed], especially in the context of evolutionary conservation.

Thank you very much for your time and consideration of this work. I look forward to learning the outcome of your deliberations.

 Yours sincerely,

 Mark Gerstein

 Albert L. Williams Professor

 of Biomedical Informatics

**Overview and Significance**

Allostery, the process by which conformational information regulates protein activity, is an essential component of protein functionality and regulation. However, it is not possible to define a protein’s allosteric behavior without first identifying the essential residues that are responsible for such behavior. The current state-of-the-art tools for identifying allosteric residues are limited in scale and scope, leaving our predictions of protein behavior fundamentally incomplete. To overcome these barriers, *we propose to develop an entirely new approach based on mathematical models of large-scale protein conformational changes to identify and predict such allosteric residues, enabling the first [please fill in as needed].* Our project will entail the objectives described below. A set of well-described conformational changes will be used as input to a biophysics-based formalism for identifying allosteric residues that can act as surface cavities or information flow bottlenecks. We will then use this formalism as a framework to develop a software tool that enables users to perform this analysis on their own proteins of interest. While our tool will be fundamentally 3D-structural in nature, we are prioritizing computational efficiency in our tool’s implementation, enabling the tool to carry out structural analyses on a large scale and without high-power computing. This approach will enable the biophysics community to study general and large-scale properties of allosteric residues across the Protein Databank. We anticipate that this work will broadly inform physical biology…

**Background and Context**

Many sources of evolutionary pressure act on proteins, and these pressures are fundamental to protein function and regulation. An integrated view of these evolutionary pressures necessarily includes structural constraints such as residue packing, protein-protein interactions, and stability. However, conformational changes and the dynamic sets of configurations are critical for predicting protein evolution and function.

The energetic landscape that affects protein conformations is dynamic in nature; allosteric signals and other external changes may reconfigure and reshape the landscape, thereby shifting the relative populations of states within an ensemble of structural configurations.1 Landscape theory thus provides the conceptual underpinnings necessary to describe how the behavior and shape of proteins will change under varying conditions. A primary driving force behind the evolution of these landscapes is the need to efficiently regulate activity in response to changing cellular contexts, thereby making allostery and conformational change essential components of protein evolution.

Given the importance of allosteric regulation, several methods have been devised to identify residues that are likely to be allosteric. Many of these methods rely on direct measures of conservation2 or co-evolution3-8 or otherwise use structure to identify residues exclusively either on the surface2,9-11 or on the protein interior.12,13 Though valuable, many of these approaches are limited in terms of scale (the number of proteins that can be feasibly investigated) or the class of residues to which the method is tailored (i.e., surface or interior residues). Therefore, our predictions of protein evolution and function are fundamentally limited.

**Scientific Approach and Objectives**

Using models of protein conformational change, we propose to develop a *comprehensive mathematical framework* incorporating protein structure and dynamics that will allow us to predict allosteric residues both on the surface and in the interior of a protein. We intend to make the computational efficiency of this framework a priority, thereby enabling high-throughput analysis for large protein sets and permitting an elucidation of the general properties of allosteric residues.

Given that knowledge of protein dynamics will be so integral to this framework, we also plan to develop a pipeline for identifying alternative conformations of proteins throughout the Protein Data Bank (PDB). The identification of likely allosteric residues within this set of dynamic proteins will allow us to examine the biophysical and evolutionary features of the identified allosteric hotspots in a straightforward fashion. Finally, we will utilize this framework to generate and distribute a tool that will enable users to submit their own structures for analysis. We anticipate that this newly introduced tool will serve as a valuable addition to our existing suite of software tools for the analysis of protein motions.

***1. Identifying Allosteric Residues on the Surface of a Protein:*** To identify likely ligand binding sites (see Fig. 1A and caption), we will use a modified version of the binding leverage method as described by Mitternacht and Berezovsky.10 This approach aims to identify cavities whose occlusion would interfere with large-scale motions of the protein. Once candidate sites for each protein are generated, we will use both anisotropic network models (ANMs) and alternative crystal structures to generate models of conformational change. We will then score each site based on the degree to which deformations at the site couple to the low-frequency modes. High-scoring sites will constitute the predicted set of surface allosteric residues.

Our approach differs from previous ones in several key ways. First, our highly efficient implementation of this method will enable more exhaustive Monte Carlo searches. In contrast to other techniques, we will also use all heavy atoms in the protein when evaluating a ligand’s affinity for each location, thereby generating a more selective set of candidate sites. In addition, we will use principles from protein folding (specifically, the concept of energy gaps) in order to sensibly threshold the list of predicted sites. As a validation step, we plan on using this method in order to predict known-ligand binding sites in well-studied systems.

***2. Identifying Critical Interior Residues via Dynamic Network Analysis:*** The framework described above will capture hotspot regions at the protein surface, but the protein interior would be neglected. Allosteric residues often act within the protein interior by functioning as essential ‘bottlenecks’ in the communication between distal regions. Therefore, we plan to use principles from network theory, in conjunction with our mathematical models of conformational change, to predict allosteric residues within the protein interior.

We will model proteins of interest as networks, wherein the residues represent nodes and the edges represent the contacts between residues. In this type of model, the problem of identifying interior-critical residues is reduced to a problem of identifying nodes that participate in network bottlenecks (Fig. 1B). We will weight edges according to the correlated motions of contacting residues; a strong correlation in the motion between contacting residues implies that knowing how one residue moves better enables one to predict the motion of the other, suggesting a strong information flow between the two residues. Then, using the motion-weighted network, we will identify “communities” of nodes using the Girvan-Newman formalism.14 Finally, we will calculate the betweenness of each edge, where the betweenness of an edge is the number of shortest paths between all pairs of residues that pass through that edge, with each path representing the sum of node-node ‘distances’ assigned in the weighting scheme above. Those residues that are involved in the highest-betweenness edges between pairs of interacting communities will be identified as the interior-critical residues.

***3. STRESS (STRucturally-identified ESSential residues): A software tool for the identification of allosteric residues:*** The implementations for finding both surface- and interior-critical residues will be made available to the scientific community through a new software tool, STRESS (for STRucturally-identified ESSential residues). This tool will allow users to specify a PDB to be analyzed, and the output provided constitutes the set of identified critical residues. To magnify the impact of this work and obviate the need for long wait times, we plan to host this service on the Amazon cloud and to use an extremely efficient algorithmic implementation.

***4. High-Throughput Identification of Alternative Conformations:*** Our framework for identifying potential allosteric residues assumes that these proteins will experience pronounced conformational changes. Therefore, to better ensure that the proteins studied exhibit well-characterized distinct conformations, we will systematically identify instances of alternative conformations within the PDB (Fig. 2). Briefly, we will perform MSAs for thousands of SCOP domains, with each alignment consisting of sequence-similar and sequence-identical domains. Within each alignment, we will cluster the domains using structural similarity to determine the distinct conformational states. This will be accomplished through a combination of multidimensional scaling and a means of identifying the optimal K value in K-means clustering.15 We will then use information regarding protein motions to identify potential allosteric sites on the surface and within the interior.

**Impact**

The proposed work will have a great impact on the fields of physical biology, protein biology, and protein evolution.

**Figures**

 ****

**Fig. 2: Identifying distinct conformations**

*Top to bottom*: **a)** Identify sequence-identical proteins. **b)** For each sequence-identical group, a multiple structure alignment is performed using STAMP (the example shown here is adenylate kinase. The SCOP IDs of the cyan domains, which constitute the *holo* structure, are d3hpqb1, d3hpqa1, d2eckb1, d2ecka1, d1akeb1, and d1akea1; the IDs of the *apo* domains, in red, are d4akea1 and d4akeb1). **c)** Using the pairwise RMSD values in this structure alignment, the structures are clustered using the UPGMA algorithm, K-means with the gap statistic (δ) is performed to identify the number of distinct conformations (2 in this example; more detailed descriptions of the graph are provided in the text). **d)** The structures which exhibit multiple clusters (i.e., those with K > 1) are then taken to exhibit multiple conformations.

**Figure 1: Finding surface- and interior-allosteric residues**

*(A)* A simulated ligand probes the protein surface as a series of Monte Carlo simulations (top-left). The cavities identified may be such that occlusion with the simulated ligand strongly interferes with conformational change (top-right, in which case they are more likely to be identified as interior-critical residues, in red), or they may have little affect on conformational change (bottom). *(B)* Interior-critical residues are identified by weighting residue-residue contacts (edges) on the basis of correlated motions, and then identifying communities within the weighted network. Residues involved in the highest-betweenness interactions between communities (in red) are selected as interior-critical residues.

**References**

1. Tsai, Chung-Jung, Buyong Ma, and Ruth Nussinov. "Folding and binding cascades: shifts in energy landscapes." Proceedings of the National Academy of Sciences 96.18 (1999): 9970-9972.
2. Panjkovich, Alejandro, and Xavier Daura. "Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery." BMC structural biology 10.1 (2010): 9.
3. Lee, Jeeyeon, et al. "Surface sites for engineering allosteric control in proteins." Science 322.5900 (2008): 438-442.
4. Suel, Gürol M., et al. "Evolutionarily conserved networks of residues mediate allosteric communication in proteins." Nature Structural & Molecular Biology 10.1 (2003): 59-69.
5. S. W. Lockless, R. Ranganathan, Science 286, 295 (1999).
6. A. I. Shulman, C. Larson, D. J. Mangelsdorf, R. Ranganathan, Cell 116, 417 (2004)
7. Reynolds, Kimberly A., Richard N. McLaughlin, and Rama Ranganathan. "Hot spots for allosteric regulation on protein surfaces." Cell 147.7 (2011): 1564-1575.
8. N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan Protein sectors: evolutionary units of three-dimensional structure Cell, 138 (2009), pp. 774–786
9. Capra, John A., et al. "Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure." PLoS Comput Biol 5.12 (2009): e1000585.
10. Mitternacht, Simon, and Igor N. Berezovsky. "Binding leverage as a molecular basis for allosteric regulation." PLoS computational biology 7.9 (2011): e1002148.
11. Ming D, Wall ME: Quantifying allosteric effects in proteins. Proteins 2005, 59(4):697-707.
12. Gasper, Paul M., et al. "Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities." Proceedings of the National Academy of Sciences 109.52 (2012): 21216-21222.
13. VanWart, Adam T., et al. "Exploring residue component contributions to dynamical network models of allostery." Journal of chemical theory and computation 8.8 (2012): 2949-2961.
14. Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." Proceedings of the National Academy of Sciences 99.12 (2002): 7821-7826.
15. N Tibshirani, Robert, Guenther Walther, and Trevor Hastie. "Estimating the number of clusters in a data set via the gap statistic." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63.2 (2001): 411-423.