Yale University

MB&B
260/266 Whitney Avenue
PO Box 208114
New Haven, CT 06520-8114

Telephone:
203 432 6105
360 838 7861 (fax)
Mark.Gerstein@yale.edu
http://bioinfo.mbb.yale.edu
[[insert_date]]

Dear *[[insert party]]*,
I am writing to you with great interest in applying for a Simons Foundation Targeted Grant in the Mathematical Modeling of Living Systems. My research group has worked extensively on the analysis of large-scale protein conformational changes. This work has not only provided novel insights and valuable intuition regarding protein structure and conformational change, but it has also resulted in the release 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 are starting to leverage our expertise in this field to better understanding allosteric residues, especially in the context of conservation. Below, I have outlined our plans along these lines. I hope that you will consider this work for funding, and thank you for taking the time to review our proposal.

Yours sincerely,

Mark Gerstein
Albert L. Williams Professor
of Biomedical Informatics

Overview

Allosteric regulation is an essential component of protein functionality and regulation. However, a full understanding of a protein's allosteric behavior is not possible without first identifying the essential residues responsible for such behavior. We plan to use models of large-scale protein conformational changes in order to identify such allosteric residues. In particular, knowledge of 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. In addition, we will 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, computational efficiency will be a priority in our tool's implementation, thereby enabling the analysis of structures on a large scale. In particular, this high-throughput approach should make it possible to study general and large-scale properties of allosteric residues across the Protein Databank.

Context & Significance

A given protein is under many sources evolutionary pressure, and these pressures are fundamental to the protein's cellular function and regulation. An integrated view of these evolutionary pressures necessarily includes structural constrains such as residue packing, protein-protein interactions, and stability. However, it must also include

information relevant to the protein's conformational changes and dynamic ensemble of configurations.

The underlying energetic landscape responsible for the relative distributions of alternative conformations is dynamic in nature: allosteric signals or other external changes may reconfigure and reshape the landscape, thereby shifting the relative populations of states within an ensemble (Tsai et al, 1999). Landscape theory thus provides the conceptual underpinnings necessary to describe how proteins change behavior and shape under changing 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, as well as the role of allostery in imparting efficient functionality, several methods have been devised for the identification of likely allosteric residues. Many of these methods rely on direct measures of conservation (Panjkovich and Daura, 2012) or co-evolution (Lee et al, 2008; Suel et al, 2003; Lockless and Ranganathan, 1999; Shulman et al, 2004; Reynolds et al, 2011; Halabi et al, 2009), or they may otherwise use structure to identify residues exclusively either on the surface (Capra et al, 2009; Panjkovich and Daura, 2012; Mitternacht and Berezovsky, 2011; Ming and Wall, 2005) or focus on the protein interior (Gasper et al, 2012; VanWart et al, 2012).

Though valuable, many of these approaches may be limited in terms of scale (the numbers of proteins which may be feasibly investigated) or the class of residues to which the method is tailored (surface or interior). Using models of protein conformational change, we propose to develop a framework to predict allosteric residues at both surface and interior within one study, and we intend to make the computational efficiency of this framework a priority, thereby enabling high-throughput analysis for many proteins, and thus better enabling the elucidation of general properties of allosteric residues. This framework would directly incorporate information regarding protein structure and dynamics. Given that knowledge of protein dynamics would be so integral to this framework, we also plan to develop a pipeline for identifying alternative conformations of proteins throughout the Protein Data Bank. Once we identify likely allosteric residues within this set of dynamic proteins, we may study biophysical and evolutionary features of the identified allosteric hotspots in a straightforward fashion. Finally, our framework will be made available through a tool to enable users to submit their own structures for analysis, and 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.

Plans & Objectives

Identifying Allosteric Residues at the Surface

We will employ a modified version of the binding leverage method for identifying likely ligand binding sites (see Fig. 1A and caption), as described previously by Mitternacht and Berezovsky. The objective will essentially be the identification of cavities such that their occlusion interferes 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 general models of conformational change, and then score each site based on the degree to which deformations in the site

couple to the low-frequency modes. High-scoring sites will constitute the predicted set of surface allosteric residues.

Our approach will differ from those previously developed in several key ways. For one, our highly efficient implementation of this method will enable more exhaustive Monte Carlo searches. In addition, we will 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, we plan on using this method in order to predict known-ligand binding sites in well-studied systems.

Dynamical Network Analysis to Identify Interior-Critical Residues

The framework described above would capture hotspot regions at the protein surface, but the protein interior is neglected. Thus, we plan to use principles from network theory, in conjunction with our models of conformational change, to predict allosteric residues within the protein interior. Allosteric residues often act within the protein interior by functioning as essential 'bottlenecks' within the communication pathways between distal regions.

We will model proteins as networks, wherein residues represent nodes and edges represent contacts between residues. In this regard, 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 by 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, thereby suggesting a strong information flow between the two residues). Then, using the motion-weighted network, "communities" of nodes will be identified using the Girvan-Newman formalism (Girvan et al, 2002). Finally, the betweenness of each edge is will be calculated (the betweenness of an edge 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), and those residues that are involved in the highest-betweenness edges between pairs of interacting communities will be identified as the interior-critical residues.

Software Tool: STRESS (STRucturally-identified ESSential residues)

The implementations for finding both surface- and interior-critical residues have will made available to the scientific community through a new software tool, STRESS. Users may specify a PDB to be analyzed, and the output provided constitutes the set of identified critical residues. Obviating the need for long wait times, the algorithmic implementation of our software will be highly efficient, and we plan on hosting this service on Amazon.

High-Throughput Identification of Alternative Conformations

Pronounced conformational change will be an essential assumption that is integral to our framework for identifying potential allosteric residues. Thus, 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). In brief, 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 (Tibshirani et al, 2001). We then use information regarding protein motions to identify potential allosteric sites on the surface and within the interior.

Figures

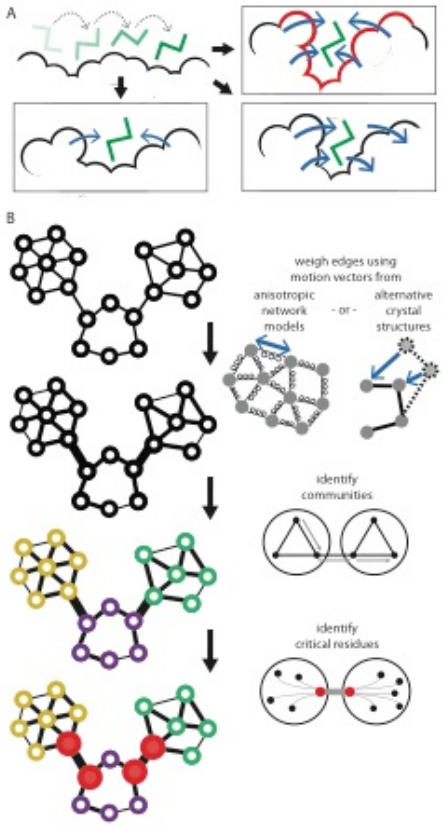

Figure 1
Finding surface- and interior-allosteric residues
(A) A simulated ligand probes the protein surface as a series of Monte Carlo simulations (top-left). The cavities identified may be such that occlusion with the simulated ligand strongly interferes with conformational change (top-right, in which 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.

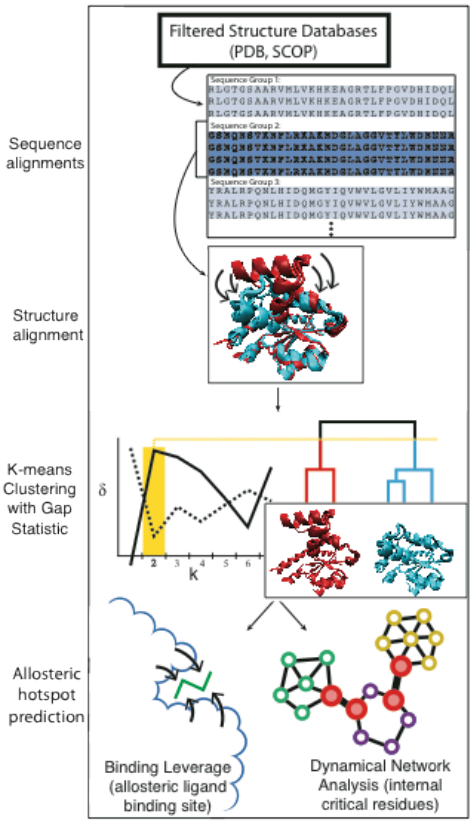

Fig. 2: Identifying distinct conformations
Top to bottom: **a)** Identify sequence-identical proteins. **b)** For each sequence-identical group, a multiple structure alignment is performed using STAMP (the example shown here is adenylate kinase. The SCOP IDs of the cyan domains, which constitute the *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.

References

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.

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.

N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan Protein sectors: evolutionary units of three-dimensional structure Cell, 138 (2009), pp. 774–786

Lee, Jeeyeon, et al. "Surface sites for engineering allosteric control in proteins." Science 322.5900 (2008): 438-442.

S. W. Lockless, R. Ranganathan, Science 286, 295 (1999).

Ming D, Wall ME: Quantifying allosteric effects in proteins. Proteins 2005, 59(4):697-707.

Mitternacht, Simon, and Igor N. Berezovsky. "Binding leverage as a molecular basis for allosteric regulation." PLoS computational biology 7.9 (2011): e1002148.

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.

Reynolds, Kimberly A., Richard N. McLaughlin, and Rama Ranganathan. "Hot spots for allosteric regulation on protein surfaces." Cell 147.7 (2011): 1564-1575.

A. I. Shulman, C. Larson, D. J. Mangelsdorf, R. Ranganathan, Cell 116, 417 (2004)

Suel, Gürol M., et al. "Evolutionarily conserved networks of residues mediate allosteric communication in proteins." Nature Structural & Molecular Biology 10.1 (2003): 59-69.

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.

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.

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.