**Overview**

 Allostery, the process by which information is transferred through a protein or complex, is an essential component of protein functionality and regulation. However, a full understanding of allosteric behavior is not possible without first identifying the essential residues responsible for such behavior. Current state-of-the-art tools for identifying allosteric residues are limited in scale (in that heavy computational demands make it difficult to study large numbers of proteins) and scope (in that many proteins are difficult to study using previously-employed experimental approaches). To overcome these barriers, we have developed a framework based on mathematical models of large-scale protein conformational changes to identify and predict such allosteric residues, thereby enabling the first evaluation of allosteric regulation on a large scale. A set of well-described conformational changes are used as input to a biophysics-based formalism for identifying allosteric residues that can act as surface cavities or information flow bottlenecks. We also introduce a software tool that enables users to perform this analysis on their own proteins of interest. While our tool is fundamentally 3D-structural in nature, we are prioritizing computational efficiency, thereby making large-scale analyses possible. This enables the biophysics community to study general properties of allosteric residues across the Protein Data Bank (PDB).

**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, this integrated view must also include information relevant to the conformational changes and dynamic ensembles of configurations.

The energetic landscapes that shape conformation are highly dynamic: allosteric signals and other external changes may reshape landscapes, thereby shifting the relative populations of states within an ensemble.1 Landscape theory thus provides the conceptual underpinnings necessary to describe how the behavior and shape of proteins 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.

Several methods have been devised to identify likely allosteric residues. 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 within the protein interior.12,13 Though valuable, existing approaches are limited in terms of scale or the types of proteins that may be studied.

The ability to identify allosteric residues on the surface and within the interior on a mass scale for many proteins would serve two purposes in particular: it would better enable the elucidation of general principles in regulation, and it would aid in developing drugs that are not limited to difficult-to-target active sites – such therapeutics may be developed to exploit allosteric regulatory sites that are distant from the active site.

**Preliminary Results**

Using models of protein conformational change, we have developed a comprehensive mathematical framework that incorporates protein structure and dynamics for predicting allosteric residues both on the surface and in the interior. Computational efficiency has been a priority in the development of this framework, thereby enabling high-throughput analysis for large protein datasets, as well as elucidating properties that are general to allosteric residues.

Given that knowledge of protein dynamics is so integral to this framework, we have also developed a pipeline for identifying alternative conformations of proteins throughout the PDB. The identification of likely allosteric residues within this set of dynamic proteins allows us to examine the biophysical and evolutionary features of the identified allosteric hotspots in a straightforward fashion. We have utilized this framework to generate and distribute a tool that enables 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. Several of the unique features include the fact that it is easy-to-use, computationally tractable, and capable of simultaneously identifying residues both at the surface and within the protein interior.

***1. Identifying Allosteric Residues on the Surface of a Protein***

To identify likely ligand binding sites, we use a modified version of the binding leverage method introduced by Mitternacht and Berezovsky.10 This approach identifies cavities whose occlusion would interfere with large-scale motions. Once candidate sites for each protein are generated, we use both anisotropic network models (ANMs) and alternative crystal structures to generate models of conformational change. We then score each site based on the degree to which deformations at the site couple to the modeled conformational changes. High-scoring sites (i.e., sites at which occlusion strongly interferes with conformational change) 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 enables more exhaustive Monte Carlo searches. In contrast to other techniques, we also use the heavy atoms of the protein when evaluating a ligand’s affinity for each location, thereby generating a more selective set of candidate sites. In addition, we 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 have implemented 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 captures hotspot regions at the protein surface, but residues in the interior may also play allosteric roles. These interior residues often act by functioning as essential ‘bottlenecks’ in the communication pathways between distal regions. Therefore, we use principles from network theory, in conjunction with our models of conformational change, to predict allosteric residues within the interior.

We model proteins as networks, wherein residues represent nodes and edges represent contacts between residues. Using this model, the problem of identifying interior-critical residues is thus reduced to a problem of identifying nodes that participate in network bottlenecks. We weigh 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 identify “communities” of nodes using the well-established Girvan-Newman formalism.14 Finally, we 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 are identified as the interior-critical residues.

***3. A software tool for the identification of allosteric residues***

The implementations for finding both surface- and interior-critical residues have been made available to the scientific community through a new software tool, STRESS (for STRucturally-identified ESSential residues). This tool allows 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 to obviate the need for long wait times, we host this service on the Amazon cloud and use an extremely efficient algorithmic implementation.

***4. High-Throughput Identification of Alternative Conformations***

Our framework for identifying potential allosteric residues assumes that these proteins undergo pronounced conformational changes. Therefore, to better ensure that the proteins studied exhibit well-characterized distinct conformations, we systematically identify instances of alternative conformations within the PDB. Briefly, we perform multiple structure alignments for thousands of structures, with each alignment consisting of sequence-identical structures. Within each alignment, we cluster the structures using structural similarity to determine the distinct conformational states. This is accomplished through a combination of multidimensional scaling and a means of identifying the optimal number of groups in K-means clustering (i.e., the “K” value).15 We then use information regarding protein motions to identify potential allosteric sites on the surface and within the interior.

**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.