ABSTRACT

It recent years, there has been a growing appreciation of the role that allostery plays in protein behavior, and it has been proposed that allostery is at play in virtually all proteins. The identification of subsets of residues that mediate allosteric signals is fundamental to gaining a fuller understanding of protein function, whereby a signal at one site affects the behavior at a remote site. The resolved structures of proteins in alternative conformations provide valuable information regarding the energetic landscapes explored, and the dynamic nature of these landscapes is reflective of the mechanisms that underlie a given protein's allosteric regulation. Herein, we describe a workflow developed to automate the identification the structures that occupy alternative energetic minima, and apply this framework to the wealth of data available in the PDB. We employ two complementary approaches in attempting to identify the key residues that likely mediate allostery. One is a refinement of the binding leverage concept introduced by (Mitternacht et al, 2011), and the second is a network-based approach employing models of conformational transitions. In sum, this work is a database-scale analysis for predicting allosteric sites on both the surface, as well as the interior residues that may join such surface sites. Notably, the complementary methods of predicting allosteric residues on the surface and interior are mechanistic in nature, in contrast to relying on conservation or biophysical properties of the individual residues. Finally, we provide a user-friendly interface to perform these predictions on submitted protein structures.

INTRODUCTION

In its early years, the PDB largely served to provide information on individual protein structures. In that only one or very few structures were available for a given protein, the data provided was static in nature. However, with the growing number of deposited structures and the concomitant saturation in the number of folds (Supp. Fig. 1), there is an increasing degree of structural redundancy, thereby expanding the opportunity from investigating macromolecular architecture to conformational heterogeneity – *a priori*, X-ray crystal structures constitute proteins in energetic wells, and may thus provide snapshots of alternative conformations. As the volume of crystal structures continues to expand, there is a growing need and opportunity to leverage this data to

identify alternative conformations, and to thus more comprehensively study their energetic landscapes.

Coincident with these trends, the mechanisms responsible for allostery and conformational change in general have been given greater attention. Identifying and discriminating between alternative conformations aids in selecting many of the proteins that are most likely to employ allostery as fundamental mechanisms of their regulation [[cite]]. Global conformational changes are often used as a primary mechanism of propagating allosteric signals through a protein. This may be the case, for instance, when the closed, inactive state of an enzyme is rendered more stable upon the binding of an inhibitor at a site that is distant from the active center.

Motivated by these ideas, we develop and apply a framework to identify instances of alternative conformations, and apply this method to the entire PDB. The repertoire of proteins that are determined to exhibit likely conformational transitions provide a rich source of raw material for investigating allosteric behavior. We emphasize that, while individual proteins have been successfully studied with methods such as molecular dynamics[[cite]] and NMR[[cite]], our protocol may be easily implemented on a database-level scale across thousands of available crystal structures in a straightforward manner, thereby more readily enabling systems-wide analyses. Notably, such a database-scale approach is much easier to exploit in studies focused on large networks of protein-protein interactions, and it may also be applied to explore the general principles and mechanisms at play in many allosteric systems once. In addition, the investigation of many proteins simultaneously also provides an means of better characterizing the large number of variants that have been shown to be deleterious through next-generation sequencing initiatives, thereby shedding light on the mechanisms at play for certain disease variants.

After identifying the conformations that occupy alternative energetic minima, we determine sites on the protein that could potentially affect the thermodynamic stability of these conformational states. This knowledge may be used to identify proteins for which drugs may be engineered, as well as instances in which sequence variation is likely to have the greatest impact by modifying the relative populations of different states. In addition to employing a mechanistic approach for identifying surface residues that may

be allosterically significant, we apply a networks-based community analysis to identify residues that may be essential for transmitting allosteric information, most of which are internal to the protein.

RESULTS

High-Throughput Identification of Structures in Distinct Energetic Wells

Although not all allosteric proteins undergo conformational change [[cite Rodgers, Nussinov]] and not all conformational changes are associated with allostery [[cite]], it is frequently the case that allosteric behavior is accompanied by substantial shifts in configurational space. The elucidation of allosteric behavior in the context of such shifts has traditionally been obtained through experiments on proteins that are limited in number, size, or both. Such studies [[cite Ranganathan, maybe also MD studies, others]] achieve high accuracy and yield valuable insights, but the limitations in scope and applicability of in-vitro studies on individual systems have made the systematic study across thousands of proteins infeasible. Working with many structures under a unified framework allows us to investigate the principles that are general to a diverse array of proteins at once.

Thus, we have constructed a high-confidence dataset of proteins in alternative conformations. A database of allosteric proteins, Allosteric Database (ASD), has been described previously [[cite]]. We checked the overlap of our predicted set of allosteric PDBs with those that appear in the ASD. Given that the total number of structures in the PDB was 82114 (as of 2012-06-15, which was around the time ASD was released and updated), and that there are 9611 PDBs in ASD (as of July 2014), ASD PDBs constitute 11.7% of the entire PDB. In total, our protein-based analysis predicted 976 allosteric PDBs. Of these, 131 also appear in ASD, relative to the expected number of matches, 114.

There are several reasons for the relatively poor overlap between our database and those described by Huang et al. We emphasize that ASD was built using literature curation rather than direct considerations of the physical properties of the proteins themselves. Many of the conformational changes are very subtle. In addition, the data in ASD is highly heterogeneous in nature, in that the structures vary considerably in terms

of resolution and experimental origin. Our focus is on a more confident set of lager-scale conformational changes with minimized noise.

To build the high-confidence dataset of conformational changes, we use a generalized approach to leverage the wealth of data in the PDB for systematically identifying proteins that occupy alternative energetic wells. It may be that a subset of these proteins do not exhibit allosteric behavior as part of their native functionality within cells, but the multiple energetic minima captured in their crystal structures may nevertheless be exploited for protein engineering [[cite C.J. Wilson, others]] or in pharmaceutical contexts [[cite]] (see proceeding discussion).

As a first step toward culling a high-confidence set of alternative conformations, we perform multiple structure alignments (MSAs) across sequence-identical domains as well as proteins, with these structures having been filtered by resolution and other metrics to ensure quality (see Methods and Figure 1 for details). We first worked with domains to probe for intra-domain conformational changes of functional significance. In addition, better structure alignments are generally possible at the domain level. The filtered dataset of domains contains 79% of all available crystal structures in the PDB (as of December 2013). PDB-wide MSAs across sequence-similar groups reveal that, in agreement with expectation, average pairwise root-mean-square deviation (RMSD) values increase at lower levels of sequence identity, as do QH values (QH, an alternative metric to RMSD, quantifies the degree to which residue-residue distances differ between two conformations, and is detailed in [[cite]] and Methods, Supp. Fig. 2).

After performing MSAs for each sequence-identical group of proteins, we use the resultant pairwise RMSD values to infer distinct conformational states. This is a non-trivial task, as considerable noise is inherent in the MSA of each protein (sources of noise may lie in either experimental limitations or the biology itself; examples include limited resolution and the fact that small differences in conformation occur within a single energetic well). Thus, in order to reduce false positives and confidently automate the identification of biologically relevant and truly distinct energetic wells, we apply a modified k-means clustering algorithm in order to assign each structure to a particular well within an MSA; structures that cluster together (i.e., exhibit low pairwise RMSD) constitute a given well (Fig. 2B).

This algorithm (termed K-means clustering with the gap statistic) identifies the ideal number of clusters (i.e., K) to describe a dataset in an automated way by comparison with a randomized null dataset. Intuitively, the parameter being optimized (i.e., the "gap") is the improvement conferred by using a given K value relative to this randomized null (a high gap value for a particular choice of K means that the dataset is well-described using K clusters). The algorithm is further detailed in methods and in (Tibsherani, 2001). Briefly, the K values obtained using this algorithm, which we take to represent the number of distinct energetic wells, is used to reduce the noise associated with limited crystallographic resolution and the potential for a protein to exhibit subtle conformational heterogeneity within a single well. The K values for MSAs, as well as the motivating conceptual framework, are summarized in Fig 2.

About 3000 different domains had a K-value of 1 (i.e., one conformation identified), whereas the K-values of close to 2000 domains exceed 1 (these exhibit multiple conformations). For proteins, close to 8000 had a K-value of 1, and about 1000 proteins had K values that exceed 1. When performing K-means clustering with the gap statistic, very similar results were obtained when clustering structures on the basis of pairwise RMSD or pairwise QH (Supp. Fig. 3), so we use RMSD in our downstream analyses.

To validate the output generated by this clustering algorithm, we manually annotated the MSAs of several well-studied canonical allosteric systems. The gap statistic performed well in discriminating crystal structures that were manually determined to constitute alternative biological states. Some of the key systems we manually annotated include proteins that have been very well-studied and characterized in the literature, such as tyrosine phosphatase (Wiesmann et al, 2004), DNA polymerase I (Xiang et al, 2006), adenylate kinase (Arora et al, 2007), Hsp ATPase (Liu et al, 2010), phosphoglycerate dehydrogenase (Grant et al, 1996), phosphofructokinase (Laurent et al, 1984), phosphotransferase (Kohl et al, 2005), and alanyl-tRNA synthetase (Dignam et al, 2011). For each of these cases, we manually determined that there were two main biological states (for example, with and without bound ligand, Supp. Fig 4). The gap statistic correctly determined that the appropriate K value for these cases was two.

Allosteric signals are often transmitted through inter-domain mechanisms. In some respects, proteins are preferable to domains in that a given protein often spans more than one domain. We confirmed the performance of the gap statistic on several high-scoring proteins, and the corresponding dendrograms (as well associated K values and fraction of times that the assigned K value appears out of 1000 simulations) is provided in the supplementary materials.

The processed output of this scheme for identifying high-confidence alternative conformations is provided as a flat text file in the Supplementary content (Supp. File 1). For each protein, this comprehensively annotated dataset includes the number of energetic wells (i.e., the identified number of clusters, or K), metrics to quantify both the confidence of the obtained K value and the assignment of proteins to different clusters (wells), as well as the corresponding PDB IDs and inter-cluster RMSDs. This dataset may thus be useful in many other research contexts focused on large-scale conformational changes present in the PDB.

Modified Binding Leverage for the Prediction of Allosteric Binding Sites

We use the high-confidence set of alternative conformations described above as the input to two complementary methods for predicting the residues that are most important in allosteric regulation and activity. In the first, binding sites on the protein surface (some of which may act as latent ligand binding sites) are identified using a modified version of the "binding leverage" framework for ligand binding site prediction developed by (Mitternacht et al, 2011, also detailed in Methods). Briefly, this method first entails using a series of Monte Carlo simulations to probe the protein surface (with the protein being represented with all heavy atoms) with an artificial ligand, thereby generating a series of candidate sites. Each candidate site is then scored on the basis of the degree to which the occlusion (with the artificial ligand) disrupts the large-scale motions of the protein (Fig. 1, bottom left). Finally, we introduce a formalism to set a cutoff for identifying the high-confidence sites from the top of the ranked list of scored sites. This approach results in finding an average of ~2 distinct binding sites per domain (Supp. Fig 5a; see Methods for the details on defining distinct sites).

In order to evaluate the effectiveness of this framework, we determined the extent to which this method captures known ligand-binding sites in a set of 12 well-studied systems for which the crystal structures of both the *holo* and *apo* states are available (Supp. Table 1).

We find that, out of the 12 canonical systems studied, we positively identify an average of 60% of the known ligand-binding sites. It has previously been shown that it is especially difficult to identify the sites in asparate transcarbamoylase (Mitternbach et al, 2011); excluding asparate transcarbamoylase from this analysis results in finding an average of 65% of known ligand-binding sites. We note that these statistics are achieved by coving an average of 15% of proteins' residues (Supp Table 2). For most proteins, selecting 15% of the residues is conservative -- more than 15% of the proteins' residues are involved in ligand binding for most proteins (Supp. Table 3).

We note that, for those sites that constitute false positives (sites which we predict to be important for allostery, but which nevertheless do not meet the thresholds needed for defining a known ligand binding sites), such sites still exhibit strong overlap with sites of biological interest – slightly reducing the thresholds needed for defining a positive hit substantially improves the fraction of true positives (Supp. Table 4). We also emphasize that our high-scoring sites which do not correspond to known biological ligand-binding sites may nevertheless correspond to latent sites (Bowman et al, 2015): even if no known biological function is assigned to such regions, the occlusion of such sites may still disrupt large-scale motions.

Such latent allosteric pockets may be useful in the context of drug development and targeting.

Residues that lie in our prioritized sites tend to be more conserved, on average, than other residues of the same protein with the same degree of burial (Fig. 3C). (Here, the degree of each residue, representing the number of other residues with which that residue interacts, is used to characterize burial). The conservation is evaluated using ConSurf scores (Glaser et al, 2003; Landau et al, 2005; Ashkenazy et al, 2010; Celniker et al 2013), and these results constitute the distribution of ConSurf scores for proteins in our entire dataset. Here, BL-critical residues had an average score of -0.131, whereas non-critical residues with the same degree distribution (i.e., same degree of burial within

the protein) had an average score of +0.059. The significance of the disparity was < 2.2e-16 (using a Wilcoxon rank sum test).

Dynamical Network Analysis

The binding leverage framework described above captures hotspot regions close to or at the surface of the protein, but the Monte Carlo search employed is *a priori* excluded from the protein interior[[cite Ranganathan et al, Nussinov et al]]. Thus, motivated by previous studies focused on individual proteins, such as tRNA synthetase (Sethi et al, 2009), essential metabolic enzymes (Manley et al, 2013), and the HIV envelope glycoprotein (Sethi et al, 2013), we apply communities-based network analyses to the protein complexes of our dataset to identify such residues, in addition to residues that may be closer to the surface.

Here, the nodes of the network represent individual residues, and edges between these nodes are drawn between residues that lie within a mutual proximity of 4.5 Angstroms. The first step in this analysis is to define the communities within the network. A given "community" refers to a group of residues that are highly inter-connected, but which have minimal edges to residues outside the community (see Methods). We have applied both an information-theory based method, termed "Infomap", (Lancichinetti et al, 2009) and the classical Girvan-Newman (GN) formalism (Girvan et al, 2002) to decompose networks of interacting protein residues into communities (see Methods). In order to recapitulate the contributions of various edges to information flow over the course of conformational changes, edges are weighted on the basis of the correlated movements using anisotropic network models, and a residue is deemed to be critical for allosteric signal transmission if it is involved in a highest-betweenness edge connecting two communities (see Methods). For instance, applying this method to threonine synthase results in the community partition and associated critical residues highlighted in Supp. Fig. 6.

Perhaps surprisingly, even though both methods achieve similar modularity (see Methods), we find that Infomap produces at least twice the number of communities relative to that of GN, and it thus generates many more critical residues (Supp. Table 5). For the canonical set of proteins, GN and Infomap generated an average of 12.0 and 36.8

communities, respectively (corresponding to an average of 44.8 and 201.4 critical residues, respectively). Thus, given that GN produces a more selective set of residues for each protein, the focus of our analyses is based on GN (corresponding results for Infomap are available in the in the Supplement).

Although the critical residues identified by GN do not always correspond to those identified by Infomap, the mean fraction of GN-identified critical residues that match Infomap-identified residues is 0.30 (the expected mean is 0.21, p-value=0.058), which further justifies our decision to focus on GN). Furthermore, we observe that obvious structural communities are detected when applying both methods (i.e., a community generated by GN is often the same as that generated by Infomap, and in other cases, a community generated by GN is often composed of sub-communities generated by Infomap).

As noted, the modularity from the network partitions generated by GN and Infomap are very similar (for the 12 canonical systems, the mean modularity for GN and Infomap is 0.73 and 0.68, respectively). Presumably, GN modularity values are consistently at least as high as those in Infomap because GN explicitly optimizes modularity in partitioning the network, whereas Infomap does not.

Residues which function as essential allosteric conduits of information in mediating signal transduction from one site of a protein to another are likely to be more conserved, on average, than residues which lie outside of such channels. Thus, as for the case with BL critical residues, as a validation of our method, we use the ConSurf server to evaluate conservation of residues identified as critical by GN, and compare these conservation scores to those of non-critical residues with the same degree (Fig. 3F). GN critical residues are generally found to be more conserved than non-critical residues with the same burial: the average ConSurf score for GN critical residues across our dataset of structures is -0.179, whereas that for non-critical residues with the same degree is -0.102 (the p-value of this disparity is 3.67e-11, using a Wilcoxon rand sum test).

HGMD & 1000 Genomes Analyses

Within our dataset of high-confidence alternative conformations, there are 176 distinct human proteins for which transcript IDs are available. Within this set of 176

distinct proteins, we identify 21 distinct proteins that are hit by known disease mutations, as collected from HGMD (Fig. 4A) [[cite]]. Many of these proteins have been studied for their important biomedical significance. Examples include hemoglobin, phenylalanine hydroxylase, p53, and Ras (a full list of the affected PDBs, along with the afflicted residues, are provided as a Supp. Files 2 and 3).

For 15 of this set of 21 proteins, the sites of HGMD mutations coincide with residues which lie in prioritized binding leverage sites. An example is Ras, shown in Fig. 4B; residues in red comprise the only prioritized binding leverage site for this complex, and those in orange are the sites of HGMD mutations). Likewise, 10 proteins have critical GN residues that overlap with sites of HGMD mutations. An example of such a system is p53, shown in Fig. 4C. The lists of proteins for which prioritized binding sites or GN residues are affected by HGMD are provided in Supp. Files 4 and 5.

In addition to capturing residues that are known to be essential in the context of disease, the residues we identify as critical are shown to be under negative selection in the context of modern-day humans. Specifically, our analyses of variants identified from The 1000 Genomes shows that GN residues occur at sites with significantly lower DAF values (Fig. 3E). Although this trend of rare alleles coinciding with critical residues also holds for the BL-identified critical residues, the trend is not significant at the level of 0.05 (Fig. 3B). The lack of significance may partly be a consequence of the limited number of proteins (44) of our dataset that are hit by 1000 Genomes SNPs. In addition, we note that the long tail extending to lower allele frequencies may be indicative of the possibility that only a subset of residues in our prioritized binding sites are essential for ligand binding (i.e., each one of our sites has 10 residues, but there may only be a small subset of these 10 which are important allosterically, thus explaining the long tail toward lower DAF values in Fig. 3B).

Using different DAF cutoffs (0.05% and 0.01%) to define rare alleles, we also examined the fraction of rare alleles (defined as the ratio of the number of low-DAF SNPs to all SNPs in a given protein) in critical residues and non-critical residues for those proteins for which at least 1 critical residue is hit by a 1000 Genomes SNP. The results for BL residues are summarized in Supp Fig. 7, and those for GN residues are summarized in Supp Fig. 8. In both cases, critical residues generally tend to be more

enriched in low-DAF variants, which is consistent with their potentially important roles in protein function.

METHODS

An overview of our pipeline is provided in Fig. 1, and we refer to this outline in the appropriate pipeline modules throughout. In brief, we perform MSAs for thousands of SCOP domains, with each alignment consisting of sequence-similar and sequence-identical domains. Within each alignment, we cluster the domains using structural similarity to determine the distinct conformational states. We then implement two complementary approaches for identifying likely allosteric residues: coarse-grained models of protein motions are used to primarily identify allosteric sites on the protein surface, and a dynamical network approach is used to primarily identify allosteric residues internal to the protein.

Database-Wide Multiple Structure Alignments

FASTA files of all SCOP domains were downloaded from the SCOP website (version 2.03) [[cite]]. In order to better ensure that large structural differences between sequence-identical or sequence-similar domains are a result of differing biological states (such as holo vs. apo, phosphorylated vs. unphosphorylated, etc.), and not an artifact of missing coordinates in X-ray crystal structures, the FASTA sequences used were those corresponding to the ATOM records of their respective PDBs. In total, this set comprises 162,517 FASTA sequences.

BLASTClust [[cite]] was downloaded from the NCBI database and used to organize these FASTA sequences into sequence-similar groups at seven levels of sequence identity (100%, 95%, 90%, 70%, 50%, 40%, and 30%). Thus, for instance, running BLASTClust with a parameter value of 100 provides a list of FASTA sequence groups such that each sequence within each group is 100% sequence identical, and in general, running BLASTClust with any given parameter value provides sequence groups such that each member within a group shares at least that specified degree of sequence identity with any other member of the same group (see top of Fig. 1).

To ensure that the X-Ray structures used in our downstream analysis are of sufficiently high quality, we removed all of those domains corresponding to PDB files with resolution values poorer than 2.8, as well as any PDB files with R-Free values poorer than 0.28. The question of how to set these quality thresholds is an important consideration, and was guided here by a combination of the thresholds conventionally used in other studies which rely on large datasets of structures [[cite Kosloff 2008, Burra 2009, others]], as well as the consideration that many interesting allosteric-related conformational changes may correlate with physical properties that sometimes render very high resolution values difficult (such as localized disorder or order-disorder transitions). As a result of applying these filters, 45,937 PDB IDs out of a total of 58,308 unique X-Ray structures (~79%) were kept for downstream analysis.

For each sequence-similar group at each of the seven levels of sequence identity, we performed multiple structure alignment (MSA) using only those domain structures that satisfy the criteria outlined above. Thus, the MSAs were generated only for those groups containing a minimum of two domains which pass the filtering criteria. The STAMP[[cite]] and MultiSeq [[cite]] plugins of VMD[[cite]] were used to generate the MSAs. Heteroatoms were removed from each domain prior to performing the alignments.

The quality of the resultant MSA for each sequence-similar group depends on the root structure used in the alignment. To obtain the optimal MSA for each group of N domains, we generated N MSAs, with each alignment using a different one of the N domains as the root structure. The best MSA generated (as measured by STAMP's "sc" score[[cite]]) was taken as the MSA for that group. Note that, in order to aid in performing the MSAs, MultiSeq was used to generate sequence alignments for each group.

Finally, for each of the N MSAs generated, MultiSeq was used calculate two measures of structural similarity between each pair of domains within a group: RMSD and QH. A fuller description of QH is provided in the Supplementary text.

For each group of sequence-similar domains, the final output of the structure alignment is a symmetric matrix representing all pairwise RMSD values (as well as a separate matrix representing all pairwise QH values) within that group. The matrices for all MSAs are then used as input to the K-means module.

We note that the pipeline above has been applied not only to SCOP domains, but also to individual proteins, with the only difference that only sequence-identical proteins were examined in this analysis.

Identifying Distinct Conformations in an Ensemble of Structures

For each MSA produced in the previous step, the corresponding matrix of pairwise RMSD values describes the degree and nature of structural heterogeneity among the crystal structures for a particular domain. The objective is to use this data in order to identify the biologically distinct conformations represented by an ensemble of structures. For a particular domain, there may be many available crystal structures. In total, these structures may actually represent only a small number of distinct biological states and conformations. For instance, there may be several crystal structures in which the domain is bound to its cognate ligand, while the remaining structures are in the apo state. Our framework for predicting the number of distinct conformational states in an ensemble of structures relies on a modified version of the K-means clustering algorithm.

A priori, performing K-means clustering assumes prior knowledge of the number of clusters (ie, "K") to describe a dataset. The purpose of K-means clustering with the gap statistic (Tibshirani et al, 2001) is to identify the optimal number of clusters intrinsic to a complex or noisy set of data points (which lie in N-dimensional space).

Given multiple resolved crystal structures for a given domain, this method (i.e., K-means with the gap statistic) estimates the number of conformational states represented in the ensemble of crystal structures (with these states presumably occupying different wells within the energetic landscape), thereby identifying proteins which are likely to undergo conformational change as part of their allosteric behavior.

As a first step toward clustering the structure ensemble represented by the RMSD matrix, it is necessary to convert this RMSD matrix (which explicitly represents only the *relationships* between distinct domains) into a form in which each domain is given its own set of coordinates. This step is necessary because the K-means algorithm acts directly on individual data points, rather than the distances between such points. Thus, we use multidimensional scaling [[ref Gower 1966 and Mardia, 1978]] to convert an N-by-N matrix (which provides all RMSD values between each pair of domains within a group of

N structures) into a set of N points, with each point representing a domain in (N-1)-dimensional space. The values of the N-1 coordinates assigned to each of these N points are such that the Euclidean distance between each pair of points are the same as the RMSD values in the original matrix. For an intuition into why N points must be mapped to (N-1)-dimensional space, consider an MSA between two structures. The RMSD between these two structures can be used to map the two domains to one-dimensional space, such that the distance between the points is the RMSD value. Similarly, an MSA of 3 domains may be mapped to 2-dimensional space in such a way that the pairwise distances are preserved; 4 domains may be mapped to 3-dimensional space, etc. The output of this multidimensional scaling is used as input to the K-means clustering with the gap statistic. We refer the reader to the work by Tibshirani et al for details governing how we perform K-means clustering with the gap statistic.

Once the optimal K value was determined for each of the N MSAs, we confirmed that these values accurately reflect the number of clusters by visual inspection of several randomly-selected MSAs, as well as several MSAs corresponding of domain groups known to constitute distinct conformations (we also visually inspected several negative controls, such as CAP, an allosteric protein which does not undergo conformational change [[ref]]). This visual inspection is carried using the RMSD values to generate dendrograms for each of the selected MSAs. The dendrograms are constructed using the hierarchical clustering algorithm built into R, `hclust` [[ref Murtagh 1985]], with UPGMA (mean values) used as the chosen agglomeration method[[ref Sokal et al, 1958]].

The next step is to assign each unique domain to its respective cluster in the alignment. Using the optimal K values assigned to each group/MSA, we carry out standard K-means clustering (Lloyd's algorithm) to determine domain-cluster assignments. For each sequence group, we perform 1000 K-means clustering simulations on the MDS coordinates, and take the most common partition generated in these simulations to assign each protein to its respective cluster.

We then select a representative domain from each of the assigned clusters. The representative member for each cluster is the member with the lowest Euclidean distance to the cluster mean, using the coordinates obtained by multidimensional scaling (see

description above). These cluster representatives are then taken as the distinct conformations for this protein, and are used for the binding leverage calculations and networks analyses (below).

Modified Binding Leverage Framework

With the objective of identifying allosteric residues (specifically those on the protein surface), we employed a modified version of the binding leverage method for predicting likely ligand binding sites (Fig. 1, bottom-left), as described previously by Mitternacht et al. This method is motivated by the observation that allosteric signals may be transmitted over large distances by a mechanism in which the allosteric ligand has a global affect on a protein's functionally important motions. For instance, introducing a bulky ligand into the site of an open pocket may disrupt large-scale motions if those motions normally entail that the pocket become completely collapsed in the apo protein. Such a modulation of the global motions may affect activity within sites that are distant from the allosteric ligand-binding site.

We refer the reader to the work by Mitternacht et al for details regarding the binding leverage method, though a general overview of the approach is given here. Hundreds or thousands of candidate allosteric sites are generated by simulations in which a simple ligand (comprising 2 to 8 atoms linked by bonds with fixed lengths but variable bond and dihedral angles) explores the protein's surface through many Monte Carlo steps. (Apo structures were used when probing protein surfaces for putative ligand binding sites). A simple square well potential was used to model the attractive and repulsive energy terms associated with the ligand's interaction with the surface. These energy terms depend only on the ligand atoms' distance to alpha carbon atoms in the protein, and they are blind to other heavy atoms or biophysical properties. Once these candidate sites have been produced, normal mode analysis is applied is generate a model of the apo protein's low-frequency motions. Each of the candidate sites is then scored based on the degree to which deformations in the site couple to the low-frequency modes; that is, those sites which are heavily deformed as a result of the normal mode fluctuations receive a high score (termed the binding leverage for that site), whereas sites which undergo minimal change over the course of a mode fluctuation receive a low binding

leverage score. The list of candidate sites is then processed to remove redundancy, and then ranked based on this score. The model stipulates that the high-scoring sites are those which are more likely to be binding sites. Using knowledge of the experimentally-determined binding sites (ie, from holo structures), the processed list of ranked sites is then used to evaluate predictive performance (see below).

Our approach and set of applications differ from those outlined by Mitternacht et al in several key ways. When running Monte Carlo simulations to probe the protein surface and generate candidate binding sites, we used all heavy atoms in the protein when evaluating a ligand's affinity for each location. By including heavy atoms in this way (ie, as oppose to using the protein's alpha carbon atoms exclusively), our hope is to generate a more realistic set of candidate ligand binding sites. Indeed, the exclusion of other heavy atoms leaves 'holes' in the protein which do not actually exist in the context of the dense topology of side chain atoms. Thus, by including all heavy atoms, we hope to reduce the number of false positive candidate sites, as well as more realistically model ligand binding affinities in general.

In the framework originally outlined by Mitternacht et al., an interaction between a ligand atom and an alpha carbon atom in the protein contributes -0.75 to the binding energy if the interaction distance is within the range of 5.5 to 8 Angstroms. Interaction distances greater than 8 Angstroms do not contribute to the binding energy, but distances in the range of 5.0 to 5.5 are repulsive, and those between 4.5 to 5.0 Angstroms are strongly repulsive (distances below 4.5 Angstroms are not permitted).

However, given the much higher density of atoms interacting with the ligand in our all-heavy atom model of each protein, it is necessary to accordingly change the energy parameters associated with the ligand's binding affinity. In particular, we varied both the ranges of favorable and unfavorable interactions, as well as the attractive and repulsive energies themselves (that is, we varied both the square well's width and depth when evaluating the ligand's affinity for a given site).

For well depths, we employed models using attractive potentials ranging from -0.05 to -0.75, including all intermediate factors of 0.05. For potential well widths, we tried performing the ligand simulations using the cutoff distances originally used by Mitternacht et al. (attractive in the range of 5.5 to 8.0 Angstroms, repulsive in the range

of 5.0 to 5.5, and strongly repulsive in the range of 4.5 to 5.0). However, these cutoffs, which were originally devised to model the ligand's affinity to the alpha carbon atom skeleton alone, were observed to be inappropriate when including all heavy atoms. Thus, we also performed the simulations using a revised set of cutoffs, with attractive interactions in the range of 3.5 to 4.5 Angstroms, repulsive interactions in the range of 3.0 to 3.5 Angstroms, and strongly repulsive interactions in the range of 2.5 to 3.0 Angstroms.

In order to identify the optimal set of parameters for defining the potential function, we determined which combination of parameters best predicts the known binding sites for several well-annotated ligand-binding proteins. This benchmark set of proteins comprised threonine synthase, phosphoribosyltransferase, tyrosine phosphatase, arginine kinase, and adenylate kinase. The structures for these five proteins were those provided as the biological assemblies associated with the PDB IDs 1E5X, 1XTT, 2HNP, 3JU5, 4AKE, respectively. Using this approach, an attractive term of -0.35 for ligand-protein atom interactions within the range of 3.5 to 4.5 Angstroms was determined to be the best overall.

The biological assembly files (as well as individual proteins and standard PDBs) for several well-annotated allosteric and ligand-binding proteins [[list]] were downloaded from the Protein Data Bank (PDB). These proteins were chosen on the basis of literature curation. Analyzed more proteins as gold standard (from several refs). Results are provided on server.

Network Analysis

Many algorithms have been devised to extract the community structure of networks. In a comprehensive study comparing different algorithms (Lancichinetti et al, 2009), an information-theory based approach (Rosvall et al, 2007), was shown to be one of the strongest. However, we found that, when applied to protein structures, this formalism tends to result in a very large number of small communities. Thus, in order to define community structure, we implemented the classical Girvan-Newman algorithm.

Under this framework, edges between residues within a protein structure are drawn between any two residues which have at least one heavy atom within a distance of

4.5 Angstroms (excluding adjacent residues in sequence, which are not considered to be in contact). Network edges are weighted on the basis of their correlated motions, with the motions provided by anisotropic network models. This weighting scheme is the same as that implemented by Sethi et al (see Sethi et al, 2009 for details), with the notable difference arising in the use of ANMs as an alternative to MD. We emphasize that, although the use of ANMs is more coarse-grained that ANMs, our use of ANMs is motivated by their much faster computational efficiency relative to MD. This added efficiency is a required feature for our database-scale analysis.

Specifically, the weight w_{ij} between residues i and j is set to $-\log(|C_{ij}|)$. After weights are assigned, the betweenness for each edge is calculated. Residues that are involved in the highest-betweenness interactions connecting pairs of interacting communities are assigned to be in the class GN-critical residues. Edge betweenness is defined as the total sum of shortest paths in which that edge is involved, with path lengths equal to the sum of edge weights (see Sethi et al, 2009 for a more detailed discussion).

1000 Genomes and HGMD Mapping & Analyses

All SNPs hitting protein-coding regions that result in amino acids changes (i.e., nonsynonymous SNPs) were collected from The 1000 Genomes Project (phase 3 release) [[cite]]. VCF files containing the annotated variants were generated using VAT [[cite]]. For nonsynonymous SNPs, the VCF files included the residue ID of the affected residue, as well as additional information (such as the corresponding allele frequency and residue type). To map the 1000 Genomes SNPs on to protein structures, FASTA files corresponding to the translated chain(s) of the respective transcript ID(s) were obtained using BioMart [[cite]]. FASTA files for each of the PDB structures associated with these transcript IDs (the PDB ID-transcript ID correspondence was also obtained using BioMart) were generated based on the ATOM records of the PDB files. For each given protein chain, BLAST was used to align the FASTA file obtained from BioMart with that generated from the PDB structure. The residue-residue correspondence obtained from these alignments was then used in order to map each SNP to specific residues within the PDB. As a quality assurance mechanism, we confirmed that the residue type reported in the VCF file matched that specified in the PDB file.

HGMD was used in order to identify any known disease-causing variants that hit the proteins in our dataset, and the SNPs in HGMD were mapped to all protein structures using the same formalism as that used in mapping 1000 Genomes variants. Given a particular set of PDBs (for example, the set of PDBs for which GN critical residues overlap with HGMD variant loci), the number of distinct proteins represented in this set was obtained by ensuring that no protein shares more than 90% sequence identity with any other protein in the set. (For instance, a set of 20 PDBs would actually constitute 17 *distinct* proteins if there exist 3 pairs of PDBs that are highly similar in sequence). Finally, we note that our set of 238 distinct HGMD proteins are those for which PDB structures are available, with the PDB structures satisfying the structure quality criteria outlined above.

DISCUSSION & CONCLUSION

Understanding allosteric signal transmission inevitably entails a consideration of the dynamic properties that generally accompany and are required for such allosteric behavior. Though a small number of examples in which allostery can occur without conformational changes have been discussed (Tsai et al, 2009; Nussinov et al, 2015), the fact that these specific systems have been highlighted underscores the important role played by conformational change in the vast majority of well-studied proteins, many of which have been investigated as a result of their significance in disease.

Molecular dynamics and NMR are some of the most common means of studying dynamic behavior. However, these methods have limitations when studying large and diverse protein datasets. Notably, molecular dynamics is computationally very expensive, and is thus impractical when studying large numbers of proteins. Like MD, NMR yields important insights, but NMR structures constitutes a relatively small fraction of the available structures in the PDB (currently about 10%).

The PDB has grown considerably since its inception. Though originally focused on finding structures for new proteins, there is now a great deal of redundancy in folds and proteins, and a concomitant greater degree of heterogeneity from a functional point of view (i.e., more models for a given protein in different biological states).

Herein, we leverage this redundancy in order to investigate protein conformational heterogeneity on a database-level scale. Motivated by the idea that large differences in shape correspond to distinct conformations that occupy different energetic wells (Fig. 2), we describe and implement a pipeline for the identification of structures in distinct conformations, using on a statistical formalism that, to our knowledge, has never previously been applied in the context of protein structures. In doing so, we integrate data from the large number of X-ray crystal structures in the PDB, and simultaneously avoid the use of computationally expensive processes. The distinct conformations culled in this analysis are manually determined to correspond to proteins known to be in distinct functional states, such as active and inactive, or holo and apo.

These different conformations are used as the raw material for the identification of residues that may be important in the context of their allosteric behavior. In the first of two complementary approaches for identifying such residues, we describe a modified version of the binding leverage method developed by Mitternacht et al. We introduce information about the heavy atoms when searching the protein surface for candidate sites in which the introduction of a ligand could strongly perturb the conformational changes of a protein, thereby finding sites that that more closely reflect cavities in the protein topology. Secondly, after the candidate sites are ranked by their ability to perturb the motions derived through anisotropic network models, we use a formalism originally used in the context of protein folding, the energy gap [[cite]], in order to define a threshold for selecting the high-confidence prioritized sites. We demonstrate that the set of high-confidence sites correspond to known ligand binding sites for a set of well-studied canonical allosteric systems. Furthermore, in the context of human polymorphism data, we show that nonsynonymous SNPs hitting these prioritized sites exhibit slightly lower allele frequencies than do SNPs that lie outside these sites.

In our second approach for finding allosteric residues, we employ a dynamical network-based analysis to search for sets of residues that may act as bottlenecks between communities in the protein structure, with these communities being defined with the GN formalism, with edge weights reflecting the dynamic properties of the protein. This network-based analysis finds residues which are both internal and in protein loops.

To evaluate the ability of this method to identify residues that may be important for allosteric behavior, we investigate their conservation in both an inter-species and intra-human genomes context. The residues identified using the dynamical network analysis are found to be conserved relative to other residues in the protein. More notably, the critical residues identified are tend to be more conserved than residues with the same degree distribution (i.e., number of neighboring) within protein structures. In addition, this greater conservation is also reflected in the genomes of modern-day humans: non-synonymous SNPs hit these critical residues with much lower frequency than do other non- synonymous SNPs hitting the same protein, suggesting that amino acid changes at these critical sites may be more deleterious than changes in other parts of the protein. We also find that several known disease SNPs, as culled from HGMD, hit the residues which we identify as being critical for allostery.

Allostery has previously been studied in the context of individual proteins. Such studies generally entail difficult and time-consuming experimental assays. There are several notable implications of our database-scale analysis. That we achieved compelling results suggests that the level of coarse graining (i.e., in x-ray crystal structures and using ANMs instead of MD) was low enough to still recapitulate biologically interesting findings. That this pipeline can be applied en masse also suggests avenues for future applications, including applications to PPIs, guiding experimental studies to prioritize residues that are candidates for allosteric behavior (cite Rama Ranganathan, others), the simultaneous characterization of many disease variants in a diverse set of proteins (HGMD), and drug development pipelines/screens in which a drug is targeted to groups of functionally related proteins (such as those related to a particular signaling cascade or functional module) rather than to specific individual structures.

FIGURE CAPTIONS

Figure 1

Pipeline for identifying distinct conformational states. *Top to bottom*: **a)** BLAST-CLUST is applied to the sequences corresponding to a filtered set of protein domains, thereby providing a large number of "sequence groups", with each group being characterized by a high degree of sequence homology. **b)** For each sequence group, a multiple structure

alignment of the domains 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 and in Fig X). **d)** The domains which exhibit multiple structural clusters (i.e., those with a $\delta > X$ and $K > 1$) are then probed for the presence of strong allosteric sites, using the complementary methods of binding leverage and dynamical network analysis (see Methods).

Figure 2

K-means clustering algorithm with the gap statistic. **a)** A schematized rendering of the k-means clustering algorithm; **b)** An example dendrogram and respective structures of a multiple-structure alignment, with similarity measured by RMSD. The example shown is for phosphotransferase, and the K-means algorithm with the gap statistic identifies K=2 different conformational states (manually determined to represent the holo and apo states of phosphotransferase); **c)**Histograms representing the k-values obtained across the database of SCOP domains and **d)** across PDB chains.

Figure 3

Conservation of predicted allosteric residues. **a)** Image of phosphfructokinase (pdb ID 3PFK), with red denoting sites with high binding leverage scores, and blue denoting sites with low scores. Known biological ligands are shown in white VDW rendering; **b)** Database-wide distributions of derived allele frequency (DAF) values of BL critical residues (red) and non-critical residues (blue); **c)** Corresponding distributions of ConSurf scores for BL critical residues (red) and non-critical residues (blue) **d)** Rendering of phosphfructokinase with GN critical residues highlighted as red spheres; **e)** Database-wide distributions of DAF values of GN critical residues (red) and non-critical residues

(blue) **f)** Corresponding distributions of ConSurf scores for GN critical residues (red) and non-critical residues (blue).

Figure 4

HGMD Analyses. **a)** Venn diagram illustrating the number of distinct proteins in various categories; **b)** Ras (PDB ID 1NVV) is an example of a protein for which HGMD locations coincide with prioritized BL sites. BL residues are shown as red spheres, and HGMD locations are in orange; **c)** p53 (PDB ID 2VUK) is an example of a protein for which HGMD locations coincide with GN residues. GN residues (red) and HGMD (orange).

REFERENCES

Arora, Karunesh, and Charles L. Brooks. "Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism." Proceedings of the National Academy of Sciences 104.47 (2007): 18496-18501.

Ashkenazy, Haim, et al. "ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids." Nucleic acids research (2010): gkq399.

Bowman, Gregory R., et al. "Discovery of multiple hidden allosteric sites by combining Markov state models and experiments." Proceedings of the National Academy of Sciences 112.9 (2015): 2734-2739.

Burra, Prasad V., et al. "Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure." Proceedings of the National Academy of Sciences 106.26 (2009): 10505-10510.

Celniker, Gershon, et al. "ConSurf: using evolutionary data to raise testable hypotheses about protein function." Israel Journal of Chemistry 53.3‐4 (2013): 199-206.

Dignam, John David, et al. "Allosteric interaction of nucleotides and tRNAala with E. coli alanyl-tRNA synthetase." Biochemistry 50.45 (2011): 9886-9900.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, et al. (2012) Ensembl 2012. Nucleic Acids Res 40: D84–90.

Fox, Naomi K., Steven E. Brenner, and John-Marc Chandonia. "SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures." Nucleic acids research 42.D1 (2014): D304-D309.

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.

Glaser, Fabian, et al. "ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information." Bioinformatics 19.1 (2003): 163-164.

Gower, J. C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53, 325–328.

Grant, Gregory A., David J. Schuller, and Leonard J. Banaszak. "A model for the regulation of D‐3‐phosphoglycerate dehydrogenase, a Vmax‐type allosteric enzyme." Protein science 5.1 (1996): 34-41.

Hubbard, Simon J., and Janet M. Thornton. "Naccess." Computer Program, Department of Biochemistry and Molecular Biology, University College London 2.1 (1993).

Kohl, Andreas, et al. "Allosteric inhibition of aminoglycoside phosphotransferase by a designed ankyrin repeat protein." Structure 13.8 (2005): 1131-1141

Kosloff, Mickey, and Rachel Kolodny. "Sequence‐similar, structure‐dissimilar protein pairs in the PDB." Proteins: Structure, Function, and Bioinformatics 71.2 (2008): 891-902.

Lancichinetti, Andrea, and Santo Fortunato. "Community detection algorithms: a comparative analysis." Physical review E 80.5 (2009): 056117.

Landau, Meytal, et al. "ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures." Nucleic acids research 33.suppl 2 (2005): W299-W302.

Laurent, M., et al. "Solution X-ray scattering studies of the yeast phosphofructokinase allosteric transition. Characterization of an ATP-induced conformation distinct in quaternary structure from the R and T states of the enzyme." Journal of Biological Chemistry 259.5 (1984): 3124-3126.

Liu, Ying, and Ivet Bahar. "Toward understanding allosteric signaling mechanisms in the ATPase domain of molecular chaperones." Pacific Symposium on Biocomputing. Vol. 15. 2010.

Manley, Gregory, Ivan Rivalta, and J. Patrick Loria. "Solution NMR and computational methods for understanding protein allostery." The Journal of Physical Chemistry B 117.11 (2013): 3063-3073.

Mardia, K.V. (1978) Some properties of classical multidimensional scaling. Communications on Statistics – Theory and Methods, A7, 1233–41.

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

Murtagh, F. (1985). “Multidimensional Clustering Algorithms”, in COMPSTAT Lectures 4. Wuerzburg: Physica-Verlag (for algorithmic details of algorithms used).

Nussinov, Ruth, and Chung-Jung Tsai. "Allostery without a conformational change? Revisiting the paradigm." Current opinion in structural biology 30 (2015): 17-24.

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, Antonio Del Sol, and Ruth Nussinov. "Allostery: absence of a change in shape does not imply that allostery is not at play." Journal of molecular biology 378.1 (2008): 1-11.

Rosvall, Martin, and Carl T. Bergstrom. "An information-theoretic framework for resolving community structure in complex networks." Proceedings of the National Academy of Sciences 104.18 (2007): 7327-7331.

Sethi, Anurag, et al. "Dynamical networks in tRNA: protein complexes." Proceedings of the National Academy of Sciences 106.16 (2009): 6620-6625.

Sethi, Anurag, et al. "A mechanistic understanding of allosteric immune escape pathways in the HIV-1 envelope glycoprotein." PLoS computational biology 9.5 (2013): e1003046.

Sokal R and Michener C (1958). "A statistical method for evaluating systematic relationships". University of Kansas Science Bulletin 38: 1409–1438.

Wiesmann, Christian, et al. "Allosteric inhibition of protein tyrosine phosphatase 1B." Nature structural & molecular biology 11.8 (2004): 730-737.

Xiang, Yun, et al. "Simulating the effect of DNA polymerase mutations on transition-state energetics and fidelity: Evaluating amino acid group contribution and allosteric coupling for ionized residues in human pol β." Biochemistry 45.23 (2006): 7036-7048.