**ABSTRACT**

The rapidly growing volume of data being produced by next-generation sequencing initiatives is enabling more in-depth analyses of protein conservation than previously possible. Deep sequencing is uncovering disease loci and protein regions under strong selective constraint, despite the fact that, in many cases, we cannot find intuitive biophysical reasons for such constraint (such as the need to engage in protein-protein interactions or to achieve a close-packed hydrophobic core). Allosteric hotspots may often provide the missing explanatory link. Here, we use models of protein conformational change to identify such allosteric residues. In particular, we predict allosteric residues that can act as surface cavities or information flow bottlenecks. We develop a software tool (stress.gersteinlab.org) that enables users to perform this analysis on their own proteins of interest. While our tool is fundamentally 3D-structural in nature, it is still computationally fast. This allows us to run it across the entire Protein Databank and evaluate large-scale properties of the predicted allosteric residues. We find that they tend to be significantly conserved across both long and short evolutionary time scales. Finally, we highlight specific examples in which these residues can help explain previously poorly understood disease-associated variants.

**INTRODUCTION**

The ability to sequence large numbers of human genomes is providing a much deeper view into protein evolution. When trying to understand the evolutionary pressures on a given protein, structural biologists now have at their disposal an unprecedented breadth of data regarding patterns of conservation, both across species and amongst humans. As such, there are greater opportunities to take a more integrated view of the context in which a protein and its residues function. This integrated view necessarily includes structural constrains such as residue packing, protein-protein interactions, and stability. However, deep sequencing is unearthing a class of conserved residues on which no obvious structural constraints appear to be acting. The missing link in understanding these regions may often be provided by considering the protein’s dynamic behavior and distinct functional states within an ensemble.

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. Conservation itself has been used, either in the context of conserved residues (Panjkovich and Daura, 2012), networks of co-evolving residues (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 local conservation in structure (Panjkovich and Daura, 2010). In related studies, both conservation and geometric-based searches for allosteric sites have been successfully applied to several systems (Capra et al, 2009). A number of methods employing support vector machines have also been described (Huang et al, 2006, Huang et al, 2013). Normal modes analysis, coupled with ligands of varying size, have been used to examine the extent to which bound ligands interfere with low-frequency motions, thereby identifying potentially important residues at the surface (Panjkovich and Daura, 2012; Mitternacht and Berezovsky, 2011; Ming and Wall, 2005).

In addition, the concept of ‘protein quakes’ has been introduced to explain local regions of proteins that are essential for conformation transitions (Miyashita et al 2003). A protein may relieve the strain of a high-energy configuration by local structural changes. Such local changes often occur at the focal point of allosteric behavior, and these regions may be identified in a number of ways, including modified normal modes analysis (Miyashita et al 2003) or time-resolved X-ray scattering (Arnlund et al, 2014).

Normal modes have also been used by the Bahar group to identify important subunits of proteins that act in a coherent manner for specific proteins (Chennubhotla and Bahar, 2006; Yang and Bahar, 2005). Rodgers et al have applied normal modes to identify key residues in CRP/FNR transcription factors (Rodgers, 2013). Molecular dynamics (MD) and network analyses have been used to identify interior residues that may function as allosteric bottlenecks (Sethi et al, 2009; Gasper et al, 2012; VanWart et al, 2012; see also reviews by Csermely et al, 2013, as well as Rousseau and Schymkowitz, 2005). In conjunction with NMR, Rivalta et al use MD and network analysis to identify important regions in imidazole glycerol phosphate synthase (Rivalta et al, 2012).

Though having provided valuable insights, many of these approaches may be limited in terms of scale (the numbers of proteins which may be feasibly investigated), computational demands, or the class of residues to which the method is tailored (surface or interior). Using models of protein conformational change, we identify both surface and interior residues that may act as essential allosteric regions in a computationally tractable manner, thereby enabling high-throughput analysis. This framework directly incorporates information regarding protein structure and dynamics, and it is applied to proteins throughout the PDB that exhibit conformational change. The relatively greater conservation of the residues identified (both across species and amongst modern-day humans) may help to elucidate many of the otherwise poorly understood regions in proteins. In a similar vein, several of our identified sites correspond to human disease loci for which no clear mechanism for pathogenesis had previously been proposed. Finally, our framework (termed STRESS, for STRucturally-identified ESSential residues) is made available through a tool to enable users to submit their own structures for analysis.

**RESULTS**

**Identifying Potential Allosteric Residues**

Allosteric residues at the surface generally play a regulatory role that is fundamentally different from that of allosteric residues within the protein interior. While surface residues may often constitute the sources or sinks of allosteric signals, interior residues act to transmit such signals. We use models of protein conformational change in an attempt to identify both classes of residues (Fig 1). Throughout, we term these potential allosteric residues at the surface and interior “surface-critical” and “interior-critical” residues, respectively. Critical residues are first identified in a set of 12 well-studied canonical systems for which both the *holo* and *apo* states are available (Supp. Table 1 and Supp. Fig. 1), and they are then identified on a large-scale across hundreds of distinct proteins.

***Identifying Surface-Critical Residues***

Allosteric ligands often act by binding to surface cavities and modulating protein conformational dynamics. The surface-critical residues, some of which may act as latent ligand binding sites and active sites, are first identified by finding cavities using Monte Carlo simulations to probe the surface with a flexible ligand (Fig. 1A, top-left). The degree to which cavity occlusion by the ligand disrupts large-scale conformational change is used to assign a score to each cavity – sites at which ligand occlusion strongly interferes with conformational change (Fig. 1A, top-right) earn high scores, whereas shallow pockets (Fig. 1A, bottom-left) or sites at which large-scale motions are largely unaffected earn lower scores (Fig. 1A, bottom-right). Further details are provided in SI Methods.

This approach is a modified version of the binding leverage framework introduced by Mitternacht and Berezovsky (Mitternacht and Berezovsky, 2011, see SI Methods). The main modifications include the use of heavy atoms in the protein during the Monte Carlo search, in addition to an automated means of thresholding the list of ranked scores to give a more selective set of candidate surface sites (see SI Methods). These modifications were implemented to provide a more selective set of sites (without them, an exceedingly large fraction of the protein surface would be captured; Supp. Fig. 2). We find that this modified approach results in an average of ~2 distinct sites per domain (Fig. 2A; see SI Methods for the details on defining distinct sites). The distribution for distinct sites withing entire complexes is given in Fig. 2B.

Within the canonical set of 12 proteins, we positively identify an average of 60% of the sites known to be directly involved in ligand or substrate binding (see Supp. Tables 2 and 3, Supp. Fig. 1, and supplementary note “Capturing Known Ligand-Binding Sites”). Some of the sites identified do not directly overlap with known binding regions, but we often find that these “false positives” nevertheless exhibit some degree of overlap with binding sites (Supp. Table 4). In addition, those surface-critical sites that do not match known binding sites may nevertheless correspond to latent allosteric regions: even if no known biological function is assigned to such regions, their occlusion may nevertheless disrupt large-scale motions.

***Dynamical Network Analysis to Identify Interior-Critical Residues***

The binding leverage framework described above captures hotspot regions at the protein surface, but the Monte Carlo search employed is *a priori* excluded from the protein interior. Allosteric residues often act within the protein interior by functioning as essential ‘bottlenecks’ within the communication pathways between distal regions. An allosteric signal transmitted from one region to another may conceivably take various alternative routes, but many of these routes can share a common set of residues. The removal of such a common set of residues can result in the loss of many or all of the available routes for allosteric signal transmission, thereby making these residues essential information flow bottlenecks.

To identify bottlenecks, the protein is first modeled as a network, wherein residues represent nodes and edges represent contacts between residues (in much the same way that the protein is modeled as a network in constructing anisotropic network models, see below). In this regard, the problem of identifying interior-critical residues is reduced to a problem of identifying nodes that participate in network bottlenecks (see Fig. 1B and SI Methods for details). Briefly, the network edges are first weighted 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. The weights are used to assign ‘distances’ between connecting nodes, with strong correlations resulting in shorter node-node distances.

Using the motion-weighted network, “communities” of nodes are identified using the Girvan-Newman formalism (Girvan et al, 2002). A community is a group of nodes such that each node within the community is highly inter-connected, but loosely connected to other nodes outside the community. Communities are thus densely inter-connected regions within proteins. As tangible examples, the community partitions and the resultant critical residues for the canonical set are given in Supp. Figs. 3 and 4.

Finally, the betweenness of each edge is 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 are identified as the interior-critical residues. These residues are essential for information flow between communities, as their removal would result in substantially longer paths between the residues of one community to those of another.

***STRESS (STRucturally-identified ESSential 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 (Supp. Fig. 5). 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 is highly efficient (Supp. Fig. 6). A typical structure takes only about 30 minutes on a 2.8GHz CPU. Running times are also minimized by using a scalable server architecture (Supp. Fig. 7). Light servers handle incoming user requests, and more powerful back-end servers, which perform the calculations, are automatically and dynamically scalable, thereby ensuring that they can handle varying levels of demand.

**High-Throughput Identification of Alternative Conformations**

Pronounced conformational change is 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 use a generalized approach to systematically identify instances of alternative conformations within the PDB. As a first step, we perform multiple structure alignments (MSAs) across sequence-identical proteins that are pre-filtered to ensure structural quality. We then use the resultant pairwise RMSD values to infer distinct conformational states (Supp. Figs. 8 and 9; see also SI Methods for details).

The distributions of the resultant numbers of conformations for domains and chains are given in Figs. 2C and 2D, respectively, and an overview of our dataset in the broader context of the entire PDB is given in Fig. 2E. Further summary statistics are provided in Supp. Fig. 10. We note that the alternative conformations identified arise in an extremely diverse set of biological contexts, including conformational transitions that accompany ligand binding, protein-protein or protein-nucleic acid interactions, post-translational modifications, changes in oxidation or oligomerization state, etc. (Supp. Fig. 11). The fully annotated dataset of conformational changes identified is provided as a resource in Supp. File 1 (see also Supp. Fig. 12).

**Evaluating the Conservation of Critical Residues with Various Metrics and Data Sources**

The large number of dynamic proteins culled throughout the PDB, coupled with the high algorithmic efficiency of our critical residue search implementation, provide a means of evaluating general, emergent properties of these residues on a large scale. In particular, we measure their conservation, as evaluated both across long (inter-species) and short (intra-human) evolutionary timescales. Using a variety of conservation metrics and sources of data, we find that both surface-critical (Figs. 3A-D) and interior-critical (Figs. 3E-H) are consistently more conserved than non-critical residues. We emphasize that the signatures of conservation identified not only provide a means of rationalizing many of the otherwise poorly-understood regions of proteins, but they also reinforce the functional importance of the residues believed to be allosteric.

***Conservation Across Species***

 When evaluating conservation across species, we find that both surface- and interior-critical residues tend to be significantly more conserved than non-critical residues (Figs. 3B and 3F, respectively). Surface-critical residues have an average conservation score (i.e., ConSurf score, see SI Methods) of -0.131, whereas non-critical residues with the same degree of burial have an average score of +0.059, demonstrating that surface-critical residues tend to be more conserved (p < 2.2e-16). Interior-critical residues exhibit a similar trend: the average conservation score for interior-critical residues and non-critical residues with the same degree of burial is -0.179 and -0.102, respectively (p=3.67e-11).

***Measures of Conservation Amongst Humans from Next-Generation Sequencing***

We may also use the large number of human genomes and exomes to investigate conservation, as many constrains may be human-specific and active in more recent evolutionary history. In this context, commonly used metrics for evaluating conservation include minor allele frequency (MAF) and derived allele frequency (DAF). Low MAF or DAF values are interpreted as signatures of deleteriousness, as purifying selection is prone to minimize the frequencies of harmful variants (see SI Methods).

We find that 1000 Genomes single-nucleotide variants (SNVs) hitting surface-critical residues tend to occur at lower DAF values (Fig. 3C; mean DAF values for surface- and non-critical sets are 9.10e-4 and 8.34e-4, respectively; p=0.309). Though not significant, the significance improves when examining the shift in the DAF distribution, as evaluated with a KS test (p= 0.159, Supp. Fig. 14a), and we emphasize the limited number of proteins (32) to be hit by 1000 Genomes SNVs (see SI Methods). The long tail extending to lower DAF values for surface-critical residues may suggest that only a subset of the residues in our prioritized binding sites is essential.

With respect to interior-critical residues, 1000 Genomes SNVs hit these residues with significantly lower DAF values than non-critical residues (Fig. 3G; mean DAF values for interior- and non-critical sets are 2.82e-4 and 3.12e-3, respectively; p=1.80e-05).

 Using MAF as a conservation metric, we performed a similar analysis using the data provided by the Exome Aggregation Consortium (Exome Aggregation Consortium (ExAC)[[this is a ref here]]). MAF distributions for surface- and non-critical residues in the same set of proteins are given in Fig. 3D (mean MAF values for surface - and non-critical sets are 4.09e-04 and 2.26e-04, respectively; p=1.49e-3). Although the mean value of the MAF distribution for surface-critical residues is slightly higher than that of non-critical residues, the median for surface-critical residues is substantially lower than that for non-critical residues. In addition, the overall shifts of these distributions also point to a trend of lower MAF values in surface-critical residues (Supp. Fig. 15A, KS test p=9.49e-2).

Interior-critical residues exhibit significantly lower MAF values than do MAF values for non-critical residues in the same set of proteins. MAF distributions for interior- and non-critical residues are given in Fig. 3H (mean MAF values for interior- and non-critical sets are 3.08e-05 and 3.27e-04, respectively; p=7.98e-09; see also Supp. Fig. 15B).

 In addition to allele frequency distributions, one may also evaluate the *fraction* of rare alleles as a metric for measuring selective pressure. This fraction is defined as the ratio of the number of low-DAF or low-MAF SNVs to all non-synonymous SNVs in a given protein (see SI Methods). A higher fraction is interpreted as a proxy to for greater conservation [[nec. to explain + cite?]]. Using different DAF cutoffs to define rarity (0.5% and 0.1%) for 1000 Genomes SNVs, both interior- and surface-critical residues harbor a higher fraction of rare alleles than do non-critical residues (Supp. Fig. 16 and Supp. Fig. 17, respectively), suggesting a greater degree of conservation in critical residues. Similar results are obtained when using MAF values for ExAC SNVs: we find that critical residues are generally more conserved than non-critical residues, and this result holds using different thresholds for defining rarity (Supp. Table 5).

***Comparisons Between Different Models of Protein Motions***

Conformational changes may be modeled using vectors connecting pairs of corresponding residues in crystal structures from alternative conformations (we term this approach “ACT”, for “absolute conformational transitions”). The crystal structures of such paired conformations may be obtained using the framework discussed above and further detailed in Methods. The protein motions may also be inferred from anisotropic network models (ANMs). ANMs entail modeling interacting residues as nodes linked by flexible springs, in a manner similar to elastic network models or normal modes analysis (Fig. 1B). ANMs are not only simple and straightforward to apply on a database scale, but unlike using alternative crystal structures, the motion vectors inferred may be generated using a single structure, and we thus use ANMs as our primary means of inferring motions.

We find that using vectors from either ACTs or ANMs give the same general results in terms of conservation, and note that our method is thus general with respect to how motion vectors are defined (see Supp. Fig. 13 and Supplemental note “Modeling Protein Motions by Directly Using Displacement Vectors from Alternative Conformations” for further details).

***Critical Residues in the Context of Human Disease Variants***

Directly related to conservation is the concept of SNV deleteriousness: changes in amino acid composition at specific loci may be more or less likely to result in disease. SIFT and PolyPhen are two tools for predicting such effects, and we evaluated these predictions for critical and non-critical residues hit by SNVs in ExAC. SNVs hitting critical residues exhibit significantly higher PolyPhen scores relative to non-critical residues, suggesting the potentially higher disease susceptibility at critical residues (Supp. Fig. 18; higher PolyPhen scores denote more damaging SNVs), though such significant disparities were not observed in SIFT scores (Supp. Fig. 19).

Using HGMD (Stenson et al 2014) and ClinVar (Landrum et al, 2014), we identify proteins with critical residues that coincide with disease-associated SNVs (Fig. 4A and Supp. Files 2 and 3). Several identified critical residues coincide with known disease loci for which the mechanism of pathogenicity is unclear unless an allosteric relationship is considered. The fibroblast growth factor receptor (FGFR) is a case-in-point (Fig. 4). SNVs in this protein have been linked to diseases that manifest in craniofacial defects. Dotted lines in Fig. 4B highlight poorly understood disease SNVs that coincide with our critical residues. The incorporation of surface- and interior-critical residues introduces an additional layer of annotation to the protein sequence, and may thus help to explain otherwise poorly understood disease-associated SNVs.

**DISCUSSION & CONCLUSIONS**

The same principles of energy landscape theory that dictate protein folding are integral to how proteins explore different conformations once they adopt their folded states. These landscapes are shaped not only by the protein sequence itself, but also by extrinsic conditions. Such external factors often regulate protein activity by introducing allosteric-induced changes, which ultimately reflect changes in the shapes and population distributions of the energetic landscape. In this regard, allostery provides an ideal platform from which to study protein behavior in the context of their energetic landscapes. To investigate allosteric regulation, and to simultaneously add an extra layer of annotation the context of its conservation patterns, an integrated framework to identify potential allosteric residues is essential. We introduce a framework to select such residues, using knowledge of conformational change.

To identify potential allosteric residues at the surface, heavy atoms are included when searching for sites at which the introduction of a ligand could strongly perturb conformational changes. Secondly, after these sites are identified, we use a formalism originally used in the context of protein folding (the energy gap (Bryngelson et al, 1994)), to define a threshold for selecting the high-confidence prioritized sites.

A dynamical network-based analysis is used to identify residues that may act as bottlenecks between communities within the protein interior. As with the identification of critical residues on the surface, information regarding conformational change is used in this network-based analysis: edges within the network of interacting residues and interacting communities are weighted on the basis of correlated motions between interacting residues.

When applied to many proteins with distinct conformational changes in the PDB, we investigate the conservation of potential allosteric residues in both inter-species and intra-human genomes contexts, and find that these residues tend to exhibit greater conservation in both cases, suggesting that amino acid changes at these critical sites are more deleterious than are changes to other residues. In addition, we identify several disease-associated variants for which plausible mechanisms had previously been unavailable, but for which allosteric mechanisms provide a plausible rationale.

Unlike the characterization of many other structural features, such as secondary structure assignment, residue burial, protein-protein interaction interfaces, disorder, and even stability, allostery inherently manifests in the context of dynamic behavior: it is only by considering protein motions and changes in these motions can a fuller understanding of allosteric regulation be realized. As such, MD and NMR are some of the most common means of studying allostery and dynamic behavior. However, these methods have limitations when studying large and diverse protein datasets. MD is computationally expensive and impractical when studying large numbers of proteins. NMR structure determination is extremely labor-intensive and better suited to certain classes of structures or dynamics. In addition, NMR structures constitute a relatively small fraction of structures currently available.

There are several notable implications of our database-scale analysis. Relative to sequence data, allostery and dynamic behavior are far more difficult to evaluate on a large scale. The framework described here enables one to evaluate dynamic behavior in a systemized and efficient way across many proteins, while simultaneously capturing residues on both the surface and within the interior. That this pipeline can be applied in a high-throughput manner enables the investigation of system-wide phenomena, such as the roles of potential allosteric hotspots in protein-protein interaction networks. Knowledge of such sites across many proteins may also be used to identify the best proteins and protein regions for which drugs should be engineered, as well as instances in which specific sequence variants are likely to have the greatest impact.

We emphasize that it is only by applying this framework over a database of a large number of proteins can one search for significant disparities in conservation between sites believed to be important in allostery and the rest of the protein. Such general trends may not be apparent when studying a small number of proteins or specific classes of proteins, but they become much more accessible when evaluating large protein datasets. To our knowledge, this is the first study in which the conservation of potential allosteric sites has been measured across a large database of proteins.

The ability to leverage our framework in a high-throughput manner also better enables one to match structural features with the high-throughput data generated through deep sequencing. Full human genomes and exomes are being sequenced at an increasing pace, thereby providing an unprecedented window into conservation patterns which can be human-specific or active over short evolutionary timescales. With such large volumes of data, these patterns increasingly serve as detailed signatures of selective constraints which may not only be missing in cross-species comparisons, but are also sometimes difficult to rationalize using static representations of protein structures alone.

We anticipate that, within the next decade, deep sequencing will enable structural biologists to study evolutionary conservation using sequenced human exomes just as routinely as cross-species alignments. Furthermore, intra-species metrics for conservation (such as those gleaned from 1000 Genomes data and ExAC) provide added value in that the confounding factors of cross-species comparisons are removed: different organisms evolve in different cellular and evolutionary contexts, and it can be difficult to decouple these different effects from one another. For instance, cross-species metrics of protein conservation entail comparisons between proteins which may be very different in structure, and which may impart very different functions in different cellular contexts. Sequence-variable regions across species may not be conserved, but nevertheless impart essential functionality. Intra-species comparisons, however, can provide a more direct and sensitive evaluation of constraint. Examples of intra-species selective constraints are particularly relevant in the context of human disease. The ubiquity of allosteric regulation as an essential feature in protein functionality and efficiency makes it well-suited to provide a conceptual framework for understanding many of the functional constraints acting on protein sequences. We believe that including information regarding likely allosteric hotspots as an added annotation to protein structures will provide a fuller understanding of conservation signatures, including those in disease contexts.

We also anticipate that our newly-developed tool (STRESS) will prove to be useful in these and related studies (stress.gersteinlab.org). It is both extremely fast and publically accessible, and as next-generation sequencing initiatives continue to provide a clearer picture of conservation at the residue level, structural biologists will increasingly find a need to explain the emergent conservation patterns. We believe that our tool will serve as a valuable tool toward meeting these needs for many proteins.

**METHODS**

An overview of finding surface- and interior-critical residues is given in Figs. 1A and 1B, respectively. Supp. Fig. 9 demonstrates an overview of our framework for identifying alternative conformations throughout the PDB, and only high-quality X-Ray structures were used in our analyses. Cross-species conservation scores are analyzed in those PDBs for which full ConSurf files are available through the ConSurf server. 1000 Genomes SNVs have been taken from the Phase 3 release, and ExAC SNVs were downloaded in May 2015. Further details on all methods are provided in SI Methods.

**ACKNOWLEDGMENTS**

…

**FIGURE CAPTIONS**

**Figure 1**

**Schematic overviews of methods for finding surface- and interior-critical 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.

**Figure 2**

**Summary statistics in database-wide analysis**

The distributions of the number of surface-critical sites per domain *(A)* and per complex *(B)*. The distributions of the number conformations (i.e., “K”) for domains *(C)* and chains *(D)*. Only proteins for which K exceeds 1 (for chains) are included in our analyzed dataset of multiple conformations. *(E)* Distinct proteins in our dataset within the context of the entire PDB. The set of distinct proteins is such that no pair shares more than 90% sequence identity.

**Figure 3**

**Conservation analyses of critical residues using multiple metrics and datasets.**

Surface- and interior-critical residues (red) for an example protein (phosphofructokinase, PDB 3PFK) are given in *(A)* and *(E)*, respectively. Distributions of cross-species conservation scores, 1000 Genomes SNV DAF values, and ExAC SNV MAF values for surface-critical and non-critical residues are given in *(B)*, *(C)*, and *(D)*, respectively. The same distributions corresponding to interior-critical residues are given in *(F)*, *(G)*, and *(H)*. Unless otherwise indicated, all p-values are based on Wilcoxon-rank sum tests. See SI Methods for details.

**Figure 4**

**Rationalizing disease-associated variants with potential allosteric residues in an example system**

*(A)* The structure shown is that of the fibroblast growth-factor receptor (FGFR), in VMD Surf rendering, with HGMD SNVs shown in orange, bound to FGF2, in ribbon rendering (PDB 1IIL). *(B)* Linear representation of structural annotation for FGFR. Dotted lines highlight loci that correspond to HGMD sites that coincide with critical residues, but for which other annotations fail to coincide. Deeply-buried residues are defined to be those that exhibit a relative solvent-exposed surface area of 5% or less, and binding site residues are defined as those for which at least one heavy atom falls within 4.5 Angstroms of any heavy atom in the binding partner (heparin-binding growth factor 2). The loci of PTM sites were taken from UniProt (accession no. P21802).

**REFERENCES**

1000 Genomes Project Consortium. "An integrated map of genetic variation from 1,092 human genomes." Nature 491.7422 (2012): 56-65.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389­3402

Arnlund, David, et al. "Visualizing a protein quake with time-resolved X-ray scattering at a free-electron laser." Nature methods 11.9 (2014): 923-926.

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.

Ashkenazy, Haim, Ron Unger, and Yossef Kliger. "Hidden conformations in protein structures." Bioinformatics 27.14 (2011): 1941-1947.

Bryngelson, Joseph D., et al. "Funnels, pathways, and the energy landscape of protein folding: a synthesis." Proteins: Structure, Function, and Bioinformatics 21.3 (1995): 167-195.

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.

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.

Chennubhotla C, Bahar I (2006) Markov propagation of allosteric effects in biomolecular systems: application to GroEL-GroES. Mol Syst Biol 2: 36.

Csermely, Peter, et al. "Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review." Pharmacology & therapeutics 138.3 (2013): 333-408.

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

Echols, Nathaniel, Duncan Milburn, and Mark Gerstein. "MolMovDB: analysis and visualization of conformational change and structural flexibility." Nucleic Acids Research 31.1 (2003): 478-482.

Exome Aggregation Consortium (ExAC), Cambridge, MA (URL: http://exac.broadinstitute.org) [May 2015]

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

Flores, Samuel, et al. "The Database of Macromolecular Motions: new features added at the decade mark." Nucleic acids research 34.suppl 1 (2006): D296-D301.

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.

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.

Gerstein, Mark, and Werner Krebs. "A database of macromolecular motions." Nucleic acids research 26.18 (1998): 4280-4290.

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.

Gunasekaran, K., Buyong Ma, and Ruth Nussinov. "Is allostery an intrinsic property of all dynamic proteins?" Proteins: Structure, Function, and Bioinformatics 57.3 (2004): 433-443.

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.

Habegger, Lukas, et al. "VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment." Bioinformatics 28.17 (2012): 2267-2269.

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

Huang, Zhimin, et al. "ASD: a comprehensive database of allosteric proteins and modulators." Nucleic acids research 39.suppl 1 (2011): D663-D669.

Huang,B. and Schroeder,M. (2006) Ligsitecsc: predicting ligand binding sites using the connolly surface and degree of conservation. BMC Struct. Biol., 6, 19.

Huang,W. et al. (2013) Allosite: a method for predicting allosteric sites. Bioinformatics, 29, 2357–2359.

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

Humphrey, W., Dalke, A. and Schulten, K., "VMD - Visual Molecular Dynamics", J. Molec. Graphics, 1996, vol. 14, pp. 33-38.

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.

Krebs, Werner G., and Mark Gerstein. "The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework." Nucleic Acids Research 28.8 (2000): 1665-1675.

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.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014 Jan 1;42(1):D980-5. doi: 10.1093/nar/gkt1113. PubMed PMID: 24234437.

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.

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

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

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

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.

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.

Miyashita, Osamu, José Nelson Onuchic, and Peter G. Wolynes. "Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins." Proceedings of the National Academy of Sciences 100.22 (2003): 12570-12575.

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

Murzin, Alexey G., et al. "SCOP: a structural classification of proteins database for the investigation of sequences and structures." Journal of molecular biology 247.4 (1995): 536-540.

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

O'Donoghue, Patrick, and Zaida Luthey-Schulten. "On the evolution of structure in aminoacyl-tRNA synthetases." Microbiology and Molecular Biology Reviews 67.4 (2003): 550-573.

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.

Panjkovich, Alejandro, and Xavier Daura. "Exploiting protein flexibility to predict the location of allosteric sites." BMC bioinformatics 13.1 (2012): 273.

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

Rivalta, Ivan, et al. "Allosteric pathways in imidazole glycerol phosphate synthase." Proceedings of the National Academy of Sciences 109.22 (2012): E1428-E1436.

Roberts, Elijah, et al. "MultiSeq: unifying sequence and structure data for evolutionary analysis." BMC bioinformatics 7.1 (2006): 382.

Rodgers, Thomas L., et al. "Modulation of global low-frequency motions underlies allosteric regulation: demonstration in CRP/FNR family transcription factors." PLoS biology 11.9 (2013): e1001651.

F. Rousseau, J. Schymkowitz A systems biology perspective on protein structural dynamics and signal transduction. Curr Opin Struct Biol, 15 (2005), pp. 23–30

Russell, Robert B., and Geoffrey J. Barton. "Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels." Proteins: Structure, Function, and Bioinformatics 14.2 (1992): 309-323.

Smedley, Damian, et al. "The BioMart community portal: an innovative alternative to large, centralized data repositories." Nucleic acids research (2015): gkv350.

Swain JF, Gierasch LM (2006) The changing landscape of protein allostery. Curr Opin Struct Biol 16: 102–108.

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.

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.

Tsai, Chung-Jung, and Ruth Nussinov. "A unified view of “how allostery works”." (2014): e1003394.

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.

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

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

Stenson et al (2014), The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 133:1-9.

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

Watson, James D., and Francis HC Crick. "Molecular structure of nucleic acids." Nature 171.4356 (1953): 737-738.

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.

Yang LW, Bahar I (2005) Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. Structure 13: 893–904.

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.