Classification (major category):
Biological Sciences

Classification (minor category):
Biophysics and Computational Biology

Title:
Identifying allosteric hotspots with dynamics: application to conservation in deep sequencing

Short title for mobile devices and RSS feeds:
Allostery and applications in deep sequencing

Authors & associated information:
Declan Clarke[a,1], Anurag Sethi[b,c,1], Shantao Li[b,d], Sushant Kumar[b,c], Richard W.F. Chang[e], Jieming Chen[b,f], and Mark Gerstein[b,c,d,2]

[a]Department of Chemistry, Yale University, 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520 USA
[b]Program in Computational Biology and Bioinformatics, Yale University, 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA
[c]Department of Molecular Biophysics and Biochemistry, Yale University, 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA
[d]Department of Computer Science, Yale University, 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA
[e]Yale College, 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA
[f]Integrated Graduate Program in Physical and Engineering Biology, Yale University, 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA

[1]D.C. and A.S. contributed equally to this work.

[2]To whom correspondence should be addressed:
Email: pi@gersteinlab.org
Phone: (203) 432-6105
Addr:
MB&B
260/266 Whitney Avenue PO Box 208114
New Haven, CT 06520-8114

Keywords:
allostery
networks
functional annotation

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, and we develop a software tool (stress.molmovdb.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 and tractable. This allows us to run it across the entire Protein Data Bank and to evaluate large-scale properties of the predicted allosteric residues. We find that these tend to be significantly conserved over both long and short evolutionary time scales. Finally, we highlight specific examples in which allosteric residues can help explain previously poorly understood disease-associated variants.

SIGNIFICANCE STATEMENT

Advances in genome sequencing technology are providing the sequenced human genomes and exomes of large numbers of individuals, thereby identifying regions under evolutionary pressure. Although signs of such pressures manifest throughout the genome, the mechanisms responsible are often unclear. Allostery serves as a plausible mechanism in many cases. We take a generalized approach to this problem by using protein conformational changes to identify potential allosteric residues in large numbers of proteins, and then evaluating their conservation using various measures and sources of data, including human genomes. These residues are conserved both among humans and across species, and they may sometimes aid in interpreting disease-associated mutations. We also introduce a user-friendly software tool for implementing this method.

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 constraints 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 (1). 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 (2), networks of co-evolving residues (3-8), or local conservation in structure (9). In related studies, both conservation and geometric-based searches for allosteric sites have been successfully applied to several systems (10). A number of methods employing support vector machines have also been described (11, 12). 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 (2, 13, 14).

The concept of 'protein quakes' has been introduced to explain local regions of proteins that are essential for conformation transitions (15). A protein may relieve the strain of a high-energy configuration by local structural changes. Such local changes often occur at the focal points of allosteric regulation, and these regions may be identified in a number of ways, including modified normal modes analysis (15) or time-resolved X-ray scattering (16).

Normal modes have also been used by the Bahar group to identify important subunits that act in a coherent manner for specific proteins (17, 18). Rodgers *et al* have applied normal modes to identify key residues in CRP/FNR transcription factors (19). Molecular dynamics (MD) and network analyses have been used to identify interior residues that may function as allosteric bottlenecks (20-24). In conjunction with NMR, Rivalta *et al* use MD and network analysis to identify important regions in imidazole glycerol phosphate synthase (25).

Though having provided valuable insights, many of these approaches may be limited in terms of scale (the numbers of proteins which may feasibly be 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 (26) that exhibit conformational change. The relatively greater conservation of the residues identified (both across species and amongst 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 distinct 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 (Table S1 and Fig. S1), 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. 1*A*, 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 interfere with conformational change earn high scores (Fig. 1*A*, top-right), whereas shallow pockets (Fig. 1*A*, bottom-left) or sites at which large-scale motions are largely unaffected (Fig. 1*A*, bottom-right) earn lower scores. Further details are provided in SI Methods.

This approach is a modified version of the binding leverage framework introduced by Mitternacht and Berezovsky (14) (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 (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 (Fig. S2). We find that this modified approach results in an average of ~2 distinct sites per domain (Fig. 2*A*; see SI Methods for details on defining distinct sites). The distribution for distinct sites within entire complexes is given in Fig. 2*B*.

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 Tables S2 and S3, Fig. S1, 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 (Table S4). 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 is intended to capture 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. 1*B* 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 'effective distances' between connecting nodes, with strong correlations resulting in shorter effective node-node distances.

Using the motion-weighted network, "communities" of nodes are identified using the Girvan-Newman formalism (27). 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 Figs. S3 and S4.

Finally, the betweenness of each edge is calculated. The betweenness of an edge is defined as the number of shortest paths between all pairs of residues that pass through that edge, with each path representing the sum of effective node-node distances assigned in the weighting scheme above. Those residues that are involved in the highest-betweenness edges between pairs of interacting communities are identified as the interior-critical residues. 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.

Software Tool: 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, which may be accessed at stress.molmovdb.org (Fig. S5). 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 (Fig. S6). A typical protein of ~500 residues takes only about 30 minutes on a 2.6GHz CPU. Running times are also minimized by using a scalable server architecture that runs on the Amazon cloud (Fig. S7). A light front-end server handles 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 both efficiently and economically.

High-Throughput Identification of Alternative Conformations

Pronounced conformational change is an essential assumption within our framework for identifying potential allosteric residues. We use a generalized approach to systematically identify instances of alternative conformations within the PDB. We first 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 (Figs. S8 and S9; see also SI Methods for details).

The distributions of the resultant numbers of conformations for domains and chains are given in Figs. 2*C* and 2*D*, respectively, and an overview is given in Fig. 2*E*. Further summary statistics are provided in Fig. S10. 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 states, etc. (Fig. S11). The dataset of alternative conformations identified is provided as a resource in File S1 (see also Fig. S12).

Evaluating Conservation of Critical Residues Using Various Metrics and Sources of Data

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 properties of these residues on a large scale. In particular, we measure their conservation, as evaluated both over 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. 3*A*-*D*) and interior-critical (Figs. 3*E*-*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 with the same degree of burial (Figs. 3*B* and 3*F*, respectively). Surface-critical residue sets have a mean conservation score (i.e., ConSurf score, see SI Methods) of -0.131, whereas non-critical residue sets with the same degree of burial have a mean score of +0.059 ($p < 2.2e-16$; negative conservation scores designate stronger conservation). Interior-critical residues exhibit a similar trend: the mean 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 sequenced human genomes and exomes to investigate conservation, as many constraints 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 reduce the frequencies of harmful variants (see SI Methods).

We find that 1000 Genomes (28) single-nucleotide variants (SNVs) that hit surface-critical residues tend to occur at lower DAF values (Fig. 3*C*). Though not significant, the significance improves when examining the shift in DAF distributions, as evaluated with a KS test ($p= 0.159$, Fig. S13*A*), and we point out the limited number of proteins (thirty-two) in which 1000 Genomes SNVs hit these critical sites (see SI Methods). Furthermore, 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. However 1000 Genomes SNVs tend hit interior-critical residues with significantly lower DAF values than non-critical residues (Fig. 3*G*; see also Fig. S13*B*).

Given the relatively small number of proteins to be hit by 1000 Genomes SNVs, we also analyzed data provided by the Exome Aggregation Consortium (ExAC) (29). ExAC provides sequence data for many more individuals, and the ExAC sequencing itself is performed at much higher coverage. Thus, using MAF as a conservation metric,

we performed a similar analysis using this data. MAF distributions for surface- and non-critical residues in the same set of proteins are given in Fig. 3*D*. 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, demonstrating that the majority of proteins are such that MAF values are lower in surface- than in non-critical residues. In addition, the overall shifts of these distributions also point to a trend of lower MAF values in surface-critical residues (Fig. S14*A*, KS test p=9.49e-2).

Interior-critical residues exhibit significantly lower MAF values than do non-critical residues in the same set of proteins. MAF distributions for interior- and non-critical residues are given in Fig. 3*H* (see also Fig. S14*B*).

In addition to overall 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 (i.e., rare) non-synonymous SNVs to all non-synonymous SNVs in a given protein annotation (such as all surface-critical residues of the protein, for example; see SI Methods). A higher fraction is interpreted as a proxy for greater conservation (30). Using variable DAF (MAF) cutoffs to define rarity for 1000 Genomes (ExAC) SNVs, both surface- and interior-critical residues are shown to harbor a higher fraction of rare alleles than do non-critical residues, further suggesting a greater degree of evolutionary constraint in critical residues (See Figs. S15 and S16 for 1000 Genomes and ExAC data, respectively).

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 (further details in SI Methods). The protein motions may also be inferred from anisotropic network models (ANMs) (31). ANMs entail modeling interacting residues as nodes linked by flexible springs, in a manner similar to elastic network models (32, 33) or normal modes analysis (Fig. 1*B*). 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.

Using vectors from either ACTs or ANMs give the same general results in terms of the disparities in conservation between critical and non-critical residues. This method is thus general with respect to how motion vectors are defined (see Fig. S17 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 (34) and PolyPhen (35) 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 (Fig. S18), though such significant disparities were not observed in SIFT scores (Fig. S19).

Using HGMD (36) and ClinVar (37), we identify proteins with critical residues that coincide with disease-associated SNVs (Fig. 4*A* and File S2). Several critical residues coincide with known disease loci for which the mechanism of pathogenicity is otherwise unclear (File S3). The fibroblast growth factor receptor (FGFR) is a case-in-point (Fig. 4). SNVs in FGFR have been linked to craniofacial defects. Dotted lines in Fig. 4*B* highlight poorly understood disease SNVs that coincide with critical residues. In addition, we identify Y328 as a surface-critical residue, which coincides with a disease-associated SNV from HGDM, despite the lack of confident predictions of deleteriousness by several widely used tools for predicting disease-associated SNVs, including PolyPhen (35), SIFT (34), and SNPs&GO (38). Together, these results suggest that the incorporation of surface- and interior-critical residues introduces a valuable layer of annotation to the protein sequence, and may 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 in the context of 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.

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. 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 many 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 or specific classes of proteins. 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 that can be human-specific or active over short evolutionary timescales. 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 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. Cross-species metrics of protein conservation entail comparisons between proteins that may be very

different in structure and function. Sequence-variable regions across species may not be conserved, but nevertheless impart essential functionality. Intra-species comparisons, however, can often provide a more direct and sensitive evaluation of constraint. In addition, intra-species selective constraints are particularly relevant in the context of human disease. Finally, we anticipate that our newly developed software tool will prove to be of great value in enabling investigators to study allostery in diverse contexts.

METHODS

An overview of the framework for finding surface- and interior-critical residues is given in Figs. 1. Fig. S9 provides a schematic of our pipeline for identifying alternative conformations throughout the PDB. Cross-species conservation scores were analyzed in those PDBs for which full ConSurf files are available through the ConSurf server. 1000 Genomes SNVs were 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

DC acknowledges the support of the NIH Predoctoral Program in Biophysics (T32 GM008283-24). We thank Simon Mitternacht for sharing the original source code for binding leverage calculations, as well as Koon-Kiu Yan for helpful discussions and feedback. The authors would like to thank the Exome Aggregation Consortium and the groups that provided exome variant data for comparison. A full list of contributing groups can be found at http://exac.broadinstitute.org/about

REFERENCES

1. Tsai CJ, Ma B, Nussinov R (1999) Folding and binding cascades: shifts in energy landscapes. *Proc Natl Acad Sci U S A* 96(18):9970-2.
2. Panjkovich A, Daura X (2012) Exploiting protein flexibility to predict the location of allosteric sites. *BMC Bioinformatics* 13(1):273.
3. Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: evolutionary units of three-dimensional structure. *Cell* 138(4):774-86.
4. Lee J, et al. (2008) Surface sites for engineering allosteric control in proteins. *Science* 322(5900):438-42.
5. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286(5438):295-9.
6. Reynolds KA, McLaughlin RN, Ranganathan R (2011) Hot spots for allosteric regulation on protein surfaces. *Cell* 147(7):1564-75.
7. Shulman AI, Larson C, Mangelsdorf DJ, Ranganathan R (2004) Structural determinants of allosteric ligand activation in RXR heterodimers. *Cell* 116(3):417-29.
8. Süel GM, Lockless SW, Wall MA, Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol* 10(1):59-69.
9. Panjkovich A, Daura X (2010) Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery. *BMC Struct Biol* 10:9.
10. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput Biol* 5(12):e1000585.
11. Huang B, Schroeder M (2006) LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct Biol* 6:19.
12. Huang W, et al. (2013) Allosite: a method for predicting allosteric sites. *Bioinformatics* 29(18):2357-9.
13. Ming D, Wall ME (2005) Quantifying allosteric effects in proteins. *Proteins* 59(4):697-707.
14. Mitternacht S, Berezovsky IN (2011) Binding leverage as a molecular basis for allosteric regulation. *PLoS Comput Biol* 7(9):e1002148.
15. Miyashita O, Onuchic JN, Wolynes PG (2003) Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc Natl Acad Sci U S A* 100(22):12570-5.
16. Arnlund D, et al. (2014) Visualizing a protein quake with time-resolved X-ray scattering at a free-electron laser. *Nat Methods* 11(9):923-6.
17. Chennubhotla C, Bahar I (2006) Markov propagation of allosteric effects in biomolecular systems: application to GroEL-GroES. *Mol Syst Biol* 2:36.
18. Yang L-W, Bahar I (2005) Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. *Structure* 13(6):893-904.
19. Rodgers TL, et al. (2013) Modulation of global low-frequency motions underlies allosteric regulation: demonstration in CRP/FNR family transcription factors. *PLoS*

Biol 11(9):e1001651.
20. Gasper PM, Fuglestad B, Komives EA, Markwick PRL, McCammon JA (2012) Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities. *Proc Natl Acad Sci U S A* 109(52):21216-22.
21. Csermely P, Korcsmáros T, Kiss HJM, London G, Nussinov R (2013) Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol Ther* 138(3):333-408.
22. Rousseau F, Schymkowitz J (2005) A systems biology perspective on protein structural dynamics and signal transduction. *Curr Opin Struct Biol* 15(1):23-30.
23. Sethi A, Eargle J, Black AA, Luthey-Schulten Z (2009) Dynamical networks in tRNA:protein complexes. *Proc Natl Acad Sci U S A* 106(16):6620-5.
24. Vanwart AT, Eargle J, Luthey-Schulten Z, Amaro RE (2012) Exploring residue component contributions to dynamical network models of allostery. *J Chem Theory Comput* 8(8):2949-2961.
25. Rivalta I, et al. (2012) Allosteric pathways in imidazole glycerol phosphate synthase. *Proc Natl Acad Sci U S A* 109(22):E1428-36.
26. Berman HM, et al. (2000) The Protein Data Bank. *Nucleic Acids Research* 28:235-242.
27. Girvan M, Newman MEJ. (2002) Community structure in social and biological networks. *Proc Natl Acad Sci U S A* 99(12):7821-6.
28. 1000 Genomes Project Consortium, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56-65.
29. Exome Aggregation Consortium (ExAC), Cambridge, M. (2015) http://exac.broadinstitute.org.
30. Khurana E, et al. (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* 342(6154):1235587.
31. Atilgan AR, et al. (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J* 80(1):505-15.
32. Fuglebakk E, Tiwari SP, and Reuter N (2015) Comparing the intrinsic dynamics of multiple protein structures using elastic network models. *Biochim Biophys Acta* 1850(5):911-22.
33. Tirion MM (1996) Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys Rev Lett* 77(9):1905-1908.
34. Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. *Genome Res* 11(5):863-74.
35. Adzhubei IA, et al. (2010) A method and server for predicting damaging missense mutations. *Nat Methods* 7(4):248-9.
36. Stenson PD, 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):1-9.
37. Landrum MJ, et al. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 42(Database issue):D980-5.
38. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat* 30(8):1237-44.

39. Hinsen K (2000) The molecular modeling toolkit: A new approach to molecular simulations. *J Comput Chem* 21:79–85.
40. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* 21(3):167-95.
41. Lancichinetti A, Fortunato S (2009) Community detection algorithms: a comparative analysis. *Phys Rev E Stat Nonlin Soft Matter Phys* 80(5 Pt 2):056117.
42. Rosvall M, Bergstrom CT (2007) An information-theoretic framework for resolving community structure in complex networks. *Proc Natl Acad Sci U S A* 104(18):7327-31.
43. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247(4):536-40.
44. Fox NK, Brenner SE, Chandonia J-M (2014) SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res* 42(Database issue):D304-9.
45. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389-402.
46. Russell RB, Barton GJ (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins* 14(2):309-23.
47. Roberts E, Eargle J, Wright D, Luthey-Schulten Z (2006) MultiSeq: unifying sequence and structure data for evolutionary analysis. *BMC Bioinformatics* 7:382.
48. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. *J Mol Graph* 14(1):33-8, 27-8.
49. O'Donoghue P, Luthey-Schulten Z (2003) On the evolution of structure in aminoacyl-tRNA synthetases. *Microbiol Mol Biol Rev* 67(4):550-73.
50. Tibshirani RN, Walther G, Hastie T (2001) 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):411-423.
51. Murtagh F (1985) Multidimensional clustering algorithms. *Compstat Lectures, Vienna: Physika Verlag* 1
52. Sokal RR (1958) A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull* 38:1409-1438.
53. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* 38(Web Server issue):W529-33.
54. Glaser F, et al. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* 19(1):163-4.
55. Landau M, et al. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res* 33(Web Server issue):W299-302.
56. Celniker G, et al. (2013) ConSurf: using evolutionary data to raise testable hypotheses about protein function. *Israel Journal of Chemistry* 13(3□4):199-206.
57. Habegger L, et al. (2012) VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics* 28(17):2267-9.

58. Smedley D, et al. (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res* 43(W1):W589-98.
59. Tennessen JA, et al. (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337(6090):64-9.

FIGURE CAPTIONS

Fig. 1. Shown are schematic overviews of methods for finding surface- and interior-critical residues. (*A*) A simulated ligand probes the protein surface in a series of Monte Carlo simulations (top-left). The cavities identified may be such that occlusion by the ligand strongly interferes with conformational change (top-right; such a site is likely to be identified as surface-critical, in red), or they may have little affect on conformational change, as in the case of shallow pockets (bottom-left) or pockets in which large-scale motions do not drastically affect pocket volume (bottom-right). (*B*) Interior-critical residues are identified by weighting residue-residue contacts (edges) on the basis of correlated motions, and then identifying communities within the weighted network. Residues involved in the highest-betweenness interactions between communities (in red) are selected as interior-critical residues.

Fig. 2. Summary statistics on database-wide analyses are shown. The distributions of the numbers of surface-critical sites per domain and per complex are given in (*A*) and (*B*), respectively. The distributions of the number conformations (i.e., "K") for domains and chains are given in (*C*) and (*D*), respectively. Only proteins for which K exceeds 1 (for chains) are included in our dataset of multiple conformations. (*E*) Distinct proteins in our dataset within the context of high-quality X-ray structures in the PDB that we structurally aligned. A set of distinct proteins is such that no pair shares more than 90% sequence identity.

Fig. 3. Multiple metrics and datasets reveal that critical residues tend to be conserved. Surface- and interior-critical residues (red) in phosphofructokinase (PDB 3PFK) are given in (*A*) and (*E*), respectively. Distributions of cross-species conservation scores, 1000 Genomes SNV DAF averages, and ExAC SNV MAF averages for surface- and non-critical residue sets are given in (*B*), (*C*), and (*D*), respectively. The same distributions corresponding to interior- and non-critical residue sets are given in (F), (G), and (H), respectively. In (C), means for surface- and non-critical sets are 9.10e-4 and 8.34e-4, respectively (p=0.309); corresponding means in (*D*) are 4.09e-04 and 2.26e-04,

respectively (p=1.49e-3). In (*G*), means for interior- and non-critical sets are 2.82e-4 and 3.12e-3, respectively (p=1.80e-05); corresponding means in (*H*) are 3.08e-05 and 3.27e-04, respectively (p=7.98e-09). Statistics for panels (*B*) and (*F*) are given in the main text. N = 421, 32, 84, 517, 31, and 90 structures for panels B, C, D, F, G, and H, respectively. P-values are based on Wilcoxon-rank sum tests. See SI Methods for further details.

Fig. 4. Potential allosteric residues add a layer of annotation to structures in the context of disease-associated SNVs. The structure shown (*A*) 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*) A linear representation of structural annotation for FGFR. Dotted lines highlight loci which 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 P21802).

Fig. 1

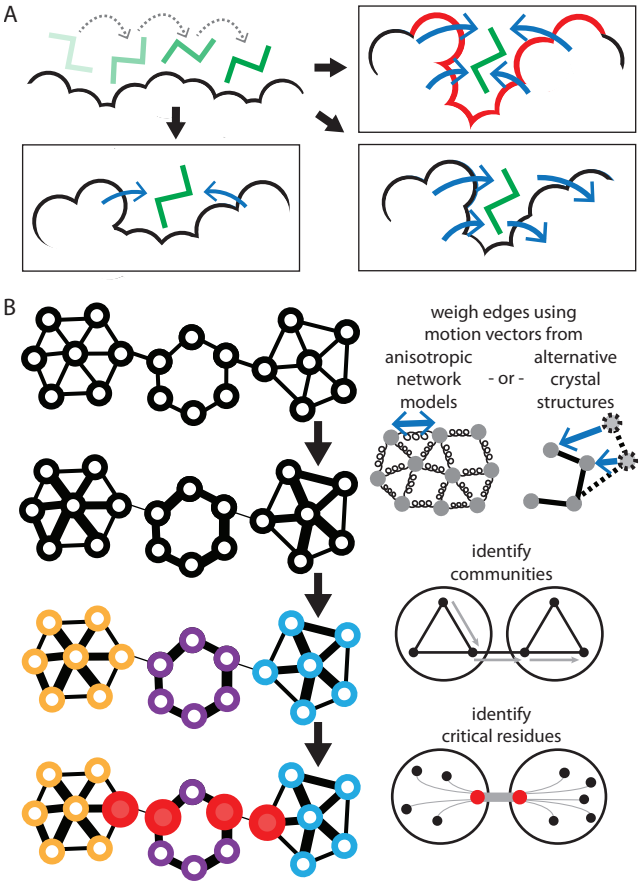

A
B
weigh edges using
motion vectors from
anisotropic
network
models
- or -
alternative
crystal
structures
identify
communities
identify
critical residues

Fig. 2

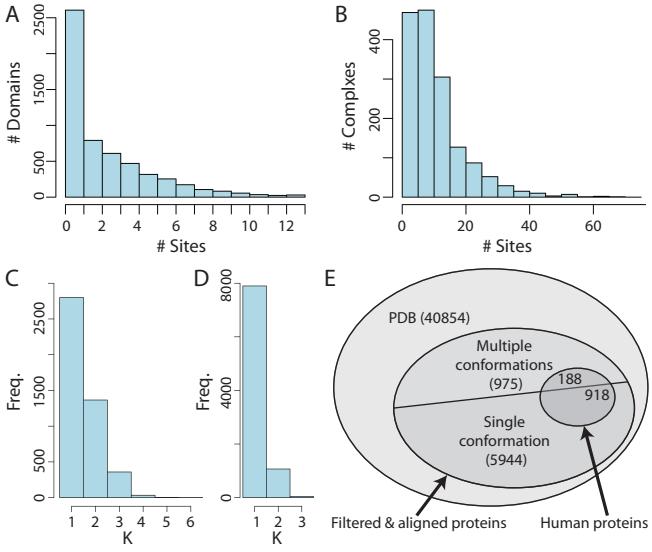

A
Domains
0
500
1500
2500
0 2 4 6 8 10 12
Sites
B
Complxes
0
200
400
0 20 40 60
Sites
C
Freq.
0
500
1500
2500
1 2 3 4 5 6
K
D
Freq.
0
4000
8000
1 2 3
K
E
PDB (40854)
Multiple
conformations
(975)
188
918
Single
conformation
(5944)
Filtered & aligned proteins
Human proteins

Fig. 3

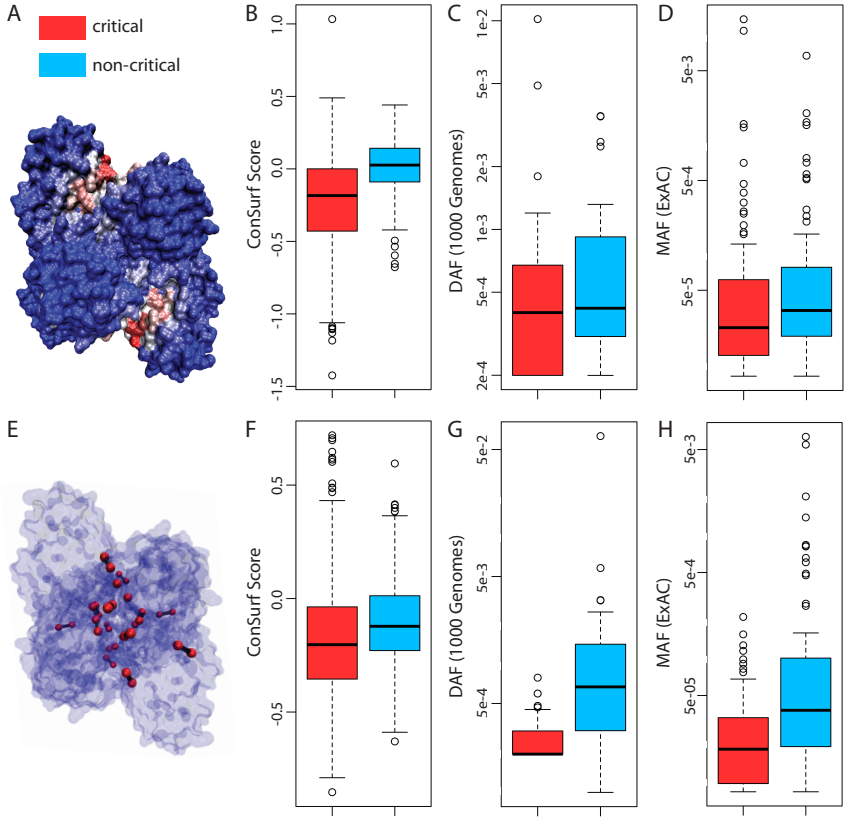
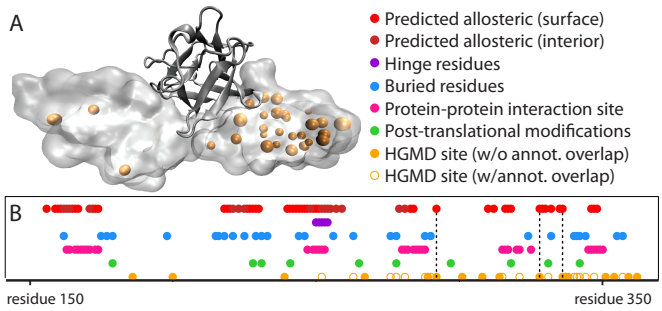

A
critical
non-critical
B
ConSurf Score
1.0
0.5
0.0
-0.5
-1.0
-1.5
C
DAF (1000 Genomes)
1e-2
5e-3
2e-3
1e-3
5e-4
2e-4
D
MAF (ExAC)
5e-3
5e-4
5e-5
E
F
ConSurf Score
0.5
0.0
-0.5
G
DAF (1000 Genomes)
5e-2
5e-3
5e-4
H
MAF (ExAC)
5e-3
5e-4
5e-05

Fig. 4


A
Predicted allosteric (surface)
Predicted allosteric (interior)
Hinge residues
Buried residues
Protein-protein interaction site
Post-translational modifications
HGMD site (w/o annot. overlap)
HGMD site (w/annot. overlap)
B
residue 150
residue 350