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 behavior, 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 interferes 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 state, 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 the large number of 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 the DAF distribution, 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 (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 MAF values for 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 conservation. 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 the formation of craniofacial defects. Dotted lines in Fig. 4*B* highlight poorly understood disease SNVs that coincide with critical residues. Notably, 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 evaluating variant pathogenicity (34, 35, 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 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.

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 of proteins 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 which 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*A* and 1*B*, respectively. 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*) 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 P21802).

Fig. 1

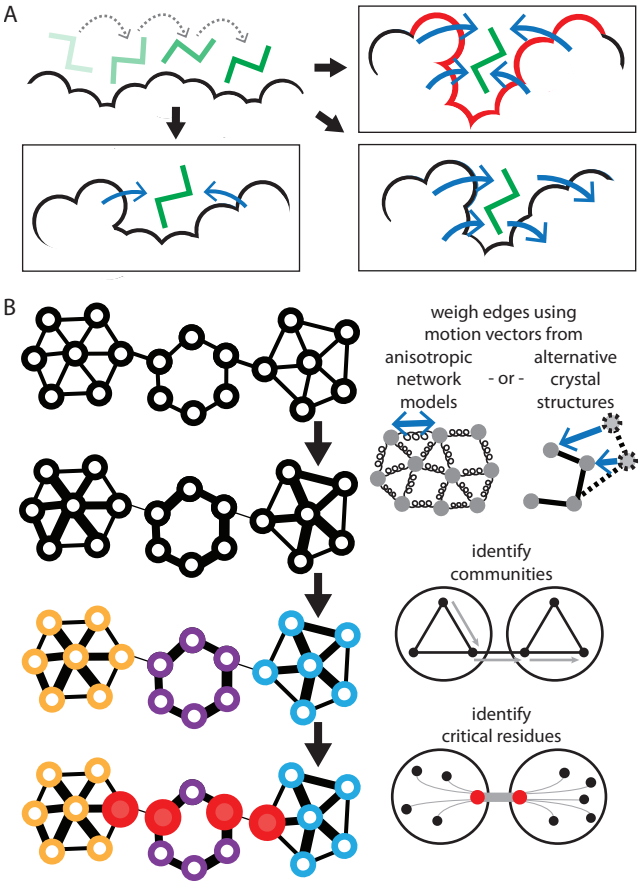

A
B
weigh edges using
motion vectors from
anisotropic
network
models
- or -
alternative
crystal
structures
identify
communities
identify
critical residues

Fig. 2

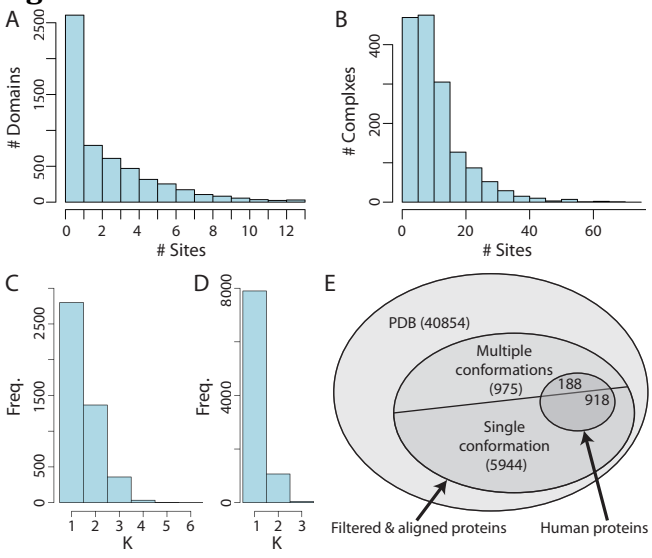

A
Domains
0
500
1500
2500
0 2 4 6 8 10 12
Sites
B
Complxes
0
200
400
0 20 40 60
Sites
C
Freq.
0
500
1500
2500
1 2 3 4 5 6
K
D
Freq.
0
4000
8000
1 2 3
K
E
PDB (40854)
Multiple
conformations
(975)
188
918
Single
conformation
(5944)
Filtered & aligned proteins
Human proteins

Fig. 3

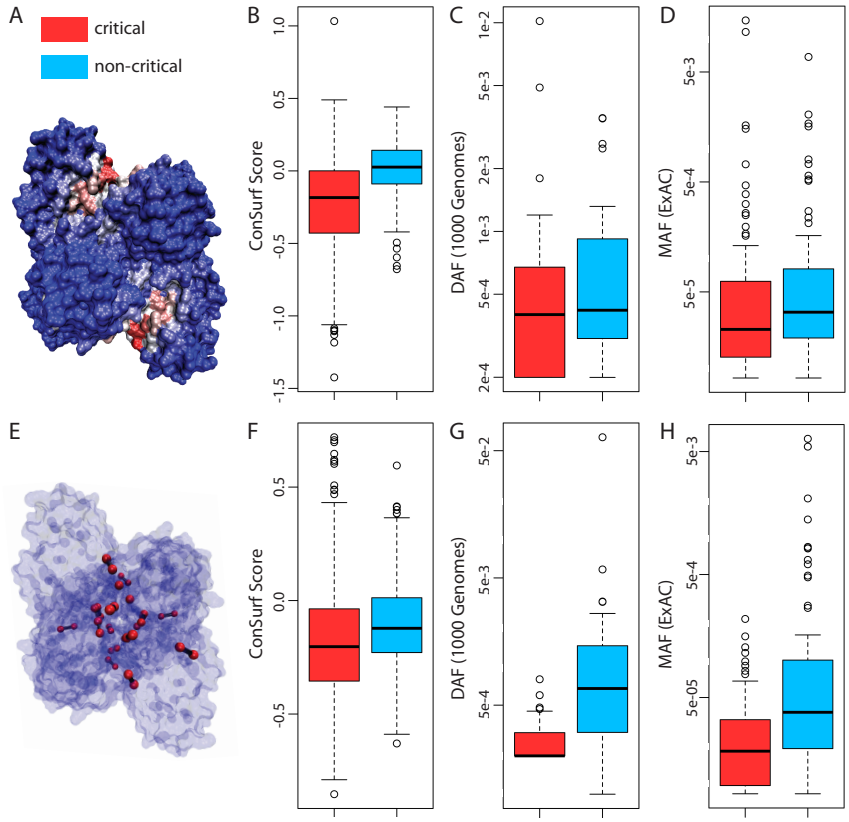
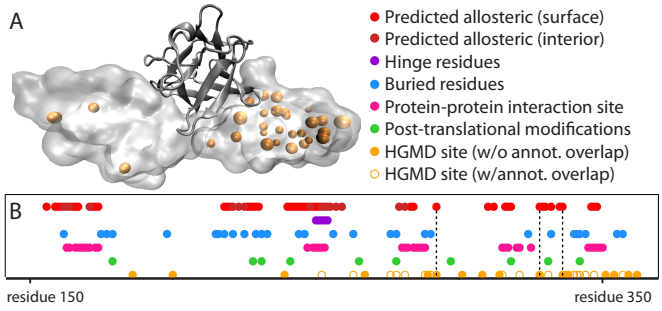

A
critical
non-critical
B
ConSurf Score
C
DAF (1000 Genomes)
D
MAF (ExAC)
E
F
ConSurf Score
G
DAF (1000 Genomes)
H
MAF (ExAC)

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

SI METHODS

Identifying Potential Allosteric Residues

Identifying Surface-Critical Residues

All biological assembly files were downloaded from the Protein Data Bank (26). With the objective of identifying potential allosteric residues on the protein surface, we employed a modified version of the binding leverage method for identifying likely ligand binding sites (Fig. 1*A*), as described previously (14). Allosteric signals may be transmitted over large distances by a mechanism in which the allosteric ligand has a global affect on a protein's functionally important motions. For instance, introducing a bulky ligand into the site of an open pocket may disrupt large-scale motions if those motions normally entail that the pocket become collapsed in the *apo* protein. Such a modulation of the global motions may affect activity within sites that are distant from the allosteric ligand-binding site.

We refer the reader to the work by Mitternacht and Berezovsky for details regarding the binding leverage method, though a general overview of the approach follows (14). Many candidate allosteric sites are generated by simulations in which a simple flexible ligand (comprising of 4 "atoms" linked by bonds of fixed length 3.8 Angstroms, but variable bond and dihedral angles) explores the protein's surface through many Monte Carlo steps. The number of Monte Carlo simulations is set to 10 times the number of residues in the protein structure, and the number of MC steps per simulation is set to 10,000 times the size of the simulation box, as measured in Angstroms. The size of this simulation box is set to twice the maximum size of the PDB along any of the x, y or z-axes. *Apo* structures were used when probing protein surfaces for putative ligand binding sites in the canonical set of proteins.

A simple square well potential (i.e., modeling hard-sphere interactions) is used to model the attractive and repulsive energy terms associated with the ligand's interaction with the surface. In the unmodified implementation of the method, these energy terms depend only on the ligand atoms' distance to *alpha carbon atoms* in the protein – other heavy atoms or biophysical properties are not considered.

Once these candidate sites are produced, normal modes analysis is applied to generate a model of the protein's low-frequency motions. To generate these modes, we use the alpha carbon atoms in building the protein's elastic networks. Using default parameters, we use the top 10 (lowest-frequency) available non-trivial modes generated using the Molecular Modeling Toolkit (MMTK) (39). Note that this exact same method for producing the modes was also used in the identification of interior-critical residues (below).

Once the modes are produced, each of the candidate sites is then scored based on the degree to which deformations in the site couple to the low-frequency modes; that is, those sites which are heavily deformed as a result of the normal mode fluctuations (Fig. 1*A*, top-right) receive a high score (termed the binding leverage for that site), whereas shallow sites with few interacting residues (Fig. 1*A*, bottom-left) or sites that undergo minimal change over the course of a mode fluctuation (Fig. 1*A*, bottom-right) receive a low binding leverage score. Strongly overlapping sites are merged, and the list is then ranked by binding leverage score. This generates a ranked list of N sites. Using knowledge of the experimentally determined binding sites (i.e., from *holo* structures), the processed list of ranked sites is then used to evaluate predictive performance (see below).

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

In the original binding leverage framework, an interaction between a ligand atom and an alpha carbon atom in the protein contributes -0.75 to the binding energy if the interaction distance is within the range of 5.5 to 8 Angstroms. Interaction distances greater than 8 Angstroms do not contribute to the binding energy, but distances in the

range of 5.0 to 5.5 are repulsive, and those between 4.5 to 5.0 Angstroms are strongly repulsive (distances below 4.5 Angstroms are not permitted). However, given the much higher density of atoms interacting with the ligand in our all-heavy atom model of each protein, it is necessary to accordingly change the energy parameters associated with the ligand's binding affinity. In particular, we varied both the ranges of favorable and unfavorable interactions, as well as the attractive and repulsive energies themselves. That is, we varied both the square well's width and depth when evaluating the ligand's affinity for a given site.

For well depths, we employed models using attractive potentials ranging from -0.05 to -0.75, including all intermediate factors of 0.05. For well widths, we tried performing the ligand simulations using the cutoff distances originally used (attractive in the range of 5.5 to 8.0 Angstroms, repulsive in the range of 5.0 to 5.5, and strongly repulsive in the range of 4.5 to 5.0). However, these cutoffs, which were originally devised to model the ligand's affinity to the alpha carbon atom skeleton alone, were observed to be inappropriate when including all heavy atoms. Thus, we also performed the simulations using a revised set of cutoffs, with attractive interactions in the range of 3.5 to 4.5 Angstroms, repulsive interactions in the range of 3.0 to 3.5 Angstroms, and strongly repulsive interactions in the range of 2.5 to 3.0 Angstroms.

In order to identify the optimal set of parameters for defining the potential function, we determined which combination of parameters best predicts the known binding sites in threonine synthase (1E5X), phosphoribosyltransferase (1XTT), tyrosine phosphatase (2HNP), arginine kinase (3JU5), and adenylate kinase (4AKE). Using this approach, an attractive term of -0.35 for ligand-protein atom interactions within the range of 3.5 to 4.5 Angstroms was determined to be the best overall.

As discussed in the main text, without applying thresholding to the list of ranked surface sites that remain after running the binding leverage calculations, a very large number of sites occupy the protein surface (Fig. S2). Thus, it is necessary to process this list. To do so, we borrow from principles in energy gap theory (40). The calculations for establishing a threshold on the number of sites are as follows:

For each of the N candidate binding sites in the pre-processed ranked list of sites, calculate $\partial BL(j)/\Delta BL$. Here, j is the j'th site to appear in the pre-processed ranked list

of sites, with this list of sites being ranked on the basis of each site's binding leverage score. $\partial BL(j)$ is defined as the difference in the binding leverage score of the j^{th} site in the ranked list and the binding leverage score in the $(j-1)^{th}$ site. Because the list of sites is organized in descending order of binding leverage scores, $\partial BL(j) \geq 0$. ΔBL is a constant that equals $\max_{binding\_leverage\_score} - \min_{binding\_leverage\_score}$ in the pre-processed ranked list of sites. ΔBL is thus the top binding leverage score that appears in this ranked list minus the bottom score. Qualitatively, a large value for $\partial BL(j)/\Delta BL$ indicates that there is a large drop in binding leverage scores in going from site (j) to site (j-1) within the pre-processed ranked list.

We then consider those sites with the highest $\partial BL/\Delta BL$ values – specifically, we consider the top 5.5% of sites in terms of their $\partial BL/\Delta BL$ values. Thus, we are considering site j if there is a very large gap in binding leverage scores between sites j and (j-1). The lowest-occurring site within *this* considered list of high $\partial BL/\Delta BL$ values demarcates a threshold beyond which we reject all lower sites within the pre-processed ranked list, leaving only the processed ranked list of sites.

We then go up from to bottom through the top of this processed ranked list of sites, and for each site, we determine the jaccard similarity with all sites above. If the jaccard similarity with any site above exceeds 0.7, then the lower site is removed from the processed ranked list. The final list obtained after performing these jaccard similarity filters is taken to represent the set of surface-critical sites on a structure.

In counting the final number of truly *distinct* surface-critical sites for any given structure, we remove redundant sites within the set of surface-critical sites obtained in the process above, as some of the sites within this set may still exhibit considerable mutual overlap. A site *i* within the list of surface-critical sites is said to be redundant if it is assigned a redundancy_score that exceeds 0.4, where redundancy_score(i) = [$residues_{site\_i} \cap \{\cup$ (set of residues in all accepted sites above site *i* in the ranked list of sites)$\}$] / [# residues in site i]. If this redundancy score is less than 0.4, then site i is included in the list of accepted sites. If it exceeds 0.4, then the site overlaps with another site on the surface, and it is thus rejected from the set of accepted distinct sites. Finally, the total number of sites in the accepted set of sites is taken as the number of distinct sites for a structure.

Capturing Known Ligand-Binding Sites

Known ligand-binding residues are those within 4.5 Angstroms of the ligand in the *holo* structure (Table S1). It has previously been shown that the sites in aspartate transcarbamoylase are especially difficult to identify (14); excluding this from this analysis results in finding an average of 65% of known biological ligand binding sites (Table S2). Note that these statistics are achieved by covering an average of 15% of proteins' residues, even though more than 15% of the proteins' residues are actually involved in ligand- or substrate-binding for most proteins (Table S3).

Dynamical Network Analysis to Identify Interior-Critical Residues

In our implementation of the Girvan-Newman framework (Fig. 1*B*), an edge between residues *i* and *j* is drawn if any heavy atom within residue *i* is located within 4.5 Angstroms of any heavy atom within residue *j*, and we exclude the trivial cases of pairs of residues that are adjacent in sequence, which are not considered to be in contact.

Network edges are then weighted on the basis of correlated motions of the interacting residues, with these motions provided by the same ANMs that had been used in the identification of surface-critical residues. We emphasize that, although ANMs are more coarse-grained that molecular dynamics, our use of ANMs is motivated by their much faster computational efficiency. This added efficiency is a required feature for our database-scale analysis. As an alternative to using ANMs, it is also possible to infer motion by simply using information regarding pairs of distinct conformations (see SI Methods subsection titled "Modeling Protein Motions by Directly Using Displacement Vectors from Alternative Conformations", below).

The edge weighting scheme is performed as follows: an "effective distance" d_{ij} for an edge between interacting residues *i* and *j* is set to $d_{ij} = -\log(|C_{ij}|)$, where C_{ij} designates the correlated motions between residue *i* and *j*:

$$C_{ij} = Cov_{ij} \; / \; \sqrt{(\langle \mathbf{r}_i^2 \rangle \langle \mathbf{r}_j^2 \rangle)}$$

where

$$Cov_{ij} = \langle \mathbf{r}_i \bullet \mathbf{r}_j \rangle$$

Here, $\mathbf{r}_i$ and $\mathbf{r}_j$ designate the vectors associated with residues *i* and *j* (respectively) under a particular mode. The brackets in $\langle \mathbf{r}_i \bullet \mathbf{r}_j \rangle$ indicate that we are taking the mean value for the dot product $\mathbf{r}_i \bullet \mathbf{r}_j$ over the 10 modes.

An example may help to clarify this. If two interacting residues exhibit a *high* degree of correlated motion, then the motion of one may tell us about the motion of the other, suggesting a strong flow of energy or information between the two residues, resulting in a *low* value for d_{ij}: a strong correlation (or a strong anti-correlation) between nodes *i* and *j* result in a value for $|C_{ij}|$ that is close to 1. This gives a low value for d_{ij} ($-\log(|C_{ij}|) \approx 0$). Thus, given a strong correlated motion, this effective distance d_{ij} between residues *i* and *j* is very short. This small d_{ij} means that any path involving this pair of residues is likewise shorter as a result, thereby more likely placing this pair of residues within a shortest path, and more likely rendering this pair a bottleneck pair. In sum, this edge-weighting scheme is such that a high correlation in motion results in a short effective distance, thereby more likely rendering this a bottleneck pair of residues.

In the opposite scenario, for interacting residues with poor correlation values ($C_{ij} \approx 0$), a large effective distance d_{ij} results, thereby making it more difficult for the pair of residues to lie within shortest paths or within the same community.

Once all connections between interacting pairs of residues are appropriately weighted and the communities are assigned using the GN algorithm with these effective distances, a residue is deemed to be critical for allosteric signal transmission (i.e., an interior-critical residue) if it is involved in the highest-betweenness edge connecting two distinct communities. A given edge's betweenness is taken to be the number of shortest paths involving that edge. Applying this method results in the community partitions and associated interior-critical residues highlighted in Figs. S3 and S4.

Decomposing Proteins into Modules Using Different Algorithms

Many algorithms have been devised to identify the community structure of networks. By this, we are referring to the problem of finding the optimal partitioning of a network into different "modules" (i.e., communties), such that each node within a module is highly connected to other nodes within the same module, and minimally conntected to other nodes in outside modules. In a comprehensive study comparing different algorithms

(41), an information theory-based approach (42), was shown to be one of the strongest. This method (termed "Infomap") effectively reduces the network community detection problem to a problem in information compression: the prominent features of the network are extracted in this compression process, giving rise to distinct modules; further details are provided in (42).

Perhaps surprisingly, even though both Infomap and GN achieve similar network modularity, Infomap (see (42)) produces at least twice the number of communities relative to that of GN when applied to protein structures, and it thus generates many more interior-critical residues (Table S5 and Fig. S20). Within the set of 12 canonical proteins, GN and Infomap generates an average of 12.0 and 36.8 communities, respectively. This corresponds to an average of 44.8 and 201.4 interior-critical residues when using GN and Infomap, respectively. Thus, given that GN produces a more selective set of residues for each protein, we use GN throughout our analyses.

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

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

STRESS (STRucturally-identified ESSential residues)

Our server has been designed to be both user-friendly and highly efficient. We use local searching supported by hashing to perform a local search in each sampling step of the Monte Carlo simulations, which takes constant time. This approach brings down the

asymptotic computational complexity by an order of magnitude, relative to a simpler implementation without optimization (Fig. S6). The time complexity of the core computation, Monte Carlo sampling, is O(|T||S|), where T and S are simulation trials and steps for each trial, respectively. After carefully profiling and optimizing for speed (with optimizations introduced through changes in the workflow, data structures, numerical arithmetic, etc.), a typical case takes ~30 minutes on a E5-2660 v3 (2.60GHz) core.

In terms of operation, our tool utilizes two types of servers: front-end servers that handle incoming HTTP requests and back-end servers that perform algorithmic calculations (Fig. S7). Communication between these two types of servers is handled by Amazon's Simple Queue Service. When our front-end servers receive a new request, they add the job to the queue and then return to requests immediately. Our back-end servers poll the queue for new jobs and run them when capacity is available. Amazon's Elastic Beanstalk offers several features that enable us to dynamically scale our web application. We use Auto Scaling to automatically adjust the number of back-end servers backing our application based on predefined conditions, such as the number of jobs in the queue and CPU utilization. Elastic Load Balancer automatically distributes incoming network traffic. This system ensures that we are able to handle varying levels of demand in a reliable and cost-effective manner. Since we may have multiple servers backing our tool simultaneously (some handling HTTP requests and some performing calculations, any of which may be terminated at any time by Auto Scaling), it is important that our servers are stateless. We thus store input and output files remotely in an S3 bucket, which is accessible to each server via RESTful conventions. The corresponding source code and README files are made available through Github (github.com/gersteinlab/STRESS).

High-Throughput Identification of Alternative Conformations

An overview of our pipeline is provided in Fig. S9. We perform MSAs for thousands of structures, with each alignment consisting of sequence-identical groups. Within each alignment, we cluster structures using RMSD to determine the distinct

conformational states. We then use information regarding protein motions to identify surface- and interior-critical residues.

Database-Wide Multiple Structure Alignments

FASTA files of all SCOP domains were downloaded from the SCOP website (version 2.03) (43, 44). We first worked with domains to probe for intra-domain conformational changes, as better alignments are generally possible at the domain level.

In order to better ensure that large structural differences between domains are a result of differing biological states (such as *holo* vs. *apo*, phosphorylated vs. unphosphorylated, etc.), and not an artifact of missing coordinates in X-ray crystal structures, the FASTA sequences used were those corresponding to the ATOM records of their respective PDBs. In total, this set comprises 162,517 FASTA sequences.

BLASTClust (45) was used to organize these FASTA sequences into sequence-similar groups at seven levels of sequence identity (100%, 95%, 90%, 70%, 50%, 40%, and 30%). Thus, for instance, running BLASTClust with a parameter value of 100 provides a list of FASTA sequence groups such that each sequence within each group is 100% sequence identical, and in general, running BLASTClust with any given parameter value provides sequence groups such that each member within a group shares at least that specified degree of sequence identity with any other member of the same group (see top of Fig. 1). Note that sequence identity values below 100% were only used to evaluate the pairwise RMSD distributions shown in Fig. S21. For all other analyses reported, all results are based on groups of structures that are 100% sequence identical.

To ensure that the X-Ray structures used in our downstream analysis are of sufficiently high quality, we removed all of those domains corresponding to PDB files with resolution values poorer than 2.8, as well as any PDB files with R-Free values poorer than 0.28. The question of how to set these quality thresholds is an important consideration, and was guided here by a combination of the thresholds conventionally used in other studies which rely on large datasets of structures, as well as the consideration that many interesting allosteric-related conformational changes may correlate with physical properties that sometimes render very high resolution values difficult (such as localized disorder or order-disorder transitions). As a result of applying

these filters, 45,937 PDB IDs out of a total of 58,308 unique X-Ray structures (~79%) were kept for downstream analysis (as of December 2013).

For each sequence-similar group at each of the seven levels of sequence identity, we performed multiple structure alignment (MSA) using only those domain structures that satisfy the criteria outlined above. Thus, the MSAs were generated only for those groups containing a minimum of two domains that pass the filtering criteria. The STAMP (46) and MultiSeq (47) plugins of VMD (48) were used to generate the MSAs. Heteroatoms were removed from each structure prior to performing the alignments.

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

Finally, for each of the N MSAs generated, MultiSeq was used calculate two measures of structural similarity between each pair of domains within a group: RMSD and Q_H. Q_H, an alternative metric to RMSD, quantifies the degree to which residue-residue distances differ between two conformations, and is detailed in (49). For each group of sequence-similar domains, the final output of the structure alignment is a symmetric matrix representing all pairwise RMSD values (as well as a separate matrix representing all pairwise Q_H values) within that group. The matrices for all MSAs are then used as input to the K-means module. PDB-wide MSAs across sequence-similar groups reveal that, in agreement with expectation, average RMSD values increase at lower levels of sequence identity (Fig. S21). Grouping structures within a multiple-structure alignment on the basis of RMSD did not change substantially when grouping structures using Q_H (Fig. S22). Thus, we use RMSD as a similarity metric throughout.

Identifying Distinct Conformations within a Multiple Structure Alignment

For each MSA produced in the previous step (using only sets of sequences that are 100% sequence-identical), the corresponding matrix of pairwise RMSD values describes the degree and nature of structural heterogeneity among the crystal structures

for a particular structure. The objective is to use this data in order to identify the biologically distinct conformations represented by an ensemble of structures. For a particular structure, there may be many available crystal structures. In total, these structures may actually represent only a small number of distinct biological states and conformations. For instance, there may be several crystal structures in which the domain is bound to its cognate ligand, while the remaining structures are in the *apo* state.

Our framework for identifying the number of distinct conformational states in an ensemble of structures relies on a modified version of the K-means clustering algorithm. This modified form of the algorithm is termed K-means clustering with the gap statistic, and it was introduced in (50).

A priori, performing K-means clustering assumes prior knowledge of the number of clusters (i.e., "K") to describe a dataset. The purpose of K-means clustering with the gap statistic is to identify the optimal number of clusters intrinsic to a complex or noisy set of data points (which lie in N-dimensional space). Given multiple resolved crystal structures for a given domain, this method estimates the number of conformational states represented in the ensemble of structures (with these states presumably occupying different wells within the energetic landscape), thereby identifying proteins which are likely to undergo conformational change as part of their functionality.

As a first step toward clustering the structure ensemble represented by an RMSD matrix, it is necessary to convert this RMSD matrix (which explicitly represents only the *relationships* between distinct structures) into a form in which each structure is given its own set of coordinates. This step is necessary because the K-means algorithm acts directly on individual data points, rather than the distances between such points. Thus, we use multidimensional scaling to convert an N-by-N matrix (which provides all RMSD values between each pair of domains within a group of N structures) into a set of N distinct points, with each point representing a domain in (N-1)-dimensional space (see below). The values of the N-1 coordinates assigned to each of these N points are such that the Euclidean distance between each pair of points are the same as the RMSD values in the original matrix.

For an intuition into why N points must be mapped to (N-1)-dimensional space, consider an MSA between two structures. The RMSD between these two structures can

be used to map the two structures to one-dimensional space, such that the distance between the points is the RMSD value. Similarly, an MSA of 3 domains may be mapped to 2-dimensional space in such a way that the pairwise distances are preserved; 4 domains may be mapped to 3-dimensional space, etc. The output of this multidimensional scaling is used as input to the K-means clustering with the gap statistic.

We refer the reader to the work by Tibshirani *et al* for details governing how we perform K-means clustering with the gap statistic, as well as more details on the theoretical justifications of this approach (50). However, an overview of the general intuition behind the formalism is provided here.

For the purpose of demonstration, assume that the data takes the form of 60 data points, with each point represented in 2D space. These are represented by the blue points in Fig. S23. Of course, our observed data in the case of MSAs may lie in N-dimensional space, in which case all Euclidean distances are just as easily calculated.

1) Start by assuming that the input data can be represented with K clusters. Perform Lloyd's algorithm (i.e., standard K-means clustering) on the dataset in order to assign each point to one of K clusters. Then, for each cluster k (which contains data points in the set C_k) measure D_k, which describes the 'density' of points within cluster k:

$$D_k = \sum_{\mathrm{x}_i \in C_k} \sum_{\mathrm{x}_j \in C_k} ||\mathrm{x}_i - \mathrm{x}_j||^2$$

2) Calculate an overall normalized score W to describe how well-clustered the resultant system has become when assigning all 60 data points to the K clusters (n_k denotes the number of points in cluster k):

$$W = \sum_{k=1}^{K} \frac{1}{2n_k} D_k$$

3) Given our observed data, how well does this number of assigned clusters K actually represent the 'true' number of clusters represented by the data, relative to a null model without any apparent clustering? To address this question, produce a null distribution of 60 randomly-distributed data points that lack any clear clustering (grey

points in Fig. S23) such that the randomly-placed points lie within the same bounding box of the observed data (blue points).

4) Repeat step (3) M times, and each time a random null distribution is produced, calculate $W_{null(K)}$ for each distribution (assuming K clusters), just as W is calculated for the observed data. Then calculate the $\text{mean}_M\{\log(W_{null(K)})\}$ for these M null distributions. Intuitively, the value $\text{mean}_M\{\log(W_{null(K)})\}$ measures how well *random* systems (with the same number of data points and within the same variable ranges as the observed data) can be described by K clusters. The M $\log(W_{null(K)})$ values produced by the null models have a standard deviation that is ultimately converted to s_k (see (50) for details):

$$s_k = \sqrt{1 + 1/B}\,\mathrm{sd}(k)$$

5) Calculate the gap statistic δ(K), given K clusters – this is a measure of how well our observed data may be described by K clusters relative to null models containing the same number of points and within the same variable ranges (i.e., within the same bounding box). Intuitively, a high value for this statistic signifies that our data is well-described using K clusters, relative to the assignment of K clusters in a randomized null distribution. Assuming K clusters, the gap statistic is given as:

$$\delta(K) = \text{mean}_M\{\log(W_{null(K)})\} - \log(W)$$

6) Obtain successive values δ(K+1), δ(K+2), δ(K+3), etc. This is done simply by incrementing the value for K and repeating the steps (1) through (5) above. Note that the optimal value of K ($K_{optimal}$, which is 3 in our demonstration case) is taken to be the first (i.e., lowest) K such that $\delta(K) >= \delta(K+1) - s_{k+1}$:

$$K_{optimal} = \{K|\ \delta(K) >= \delta(K+1) - s_{k+1}\}$$

Once the optimal K-value was determined for each MSA, we confirmed that these values accurately reflect the number of clusters by manually studying several randomly-selected MSAs, as well as several MSAs corresponding of proteins known to constitute distinct conformations. We also examined several negative controls, such as CAP, an allosteric protein that does not undergo conformational change (19).

In manually annotating the alignments, we identified a vast array well-studied canonical allosteric domains and proteins. There may be many factors driving conformational change, and those cases for which the change is induced by the binding to a simple ligand (i.e., a simple consideration of *apo* or *holo* states) constitute only a very small subset of the conformational shifts observed in the PDB (Fig. S11). The gap statistic performed well in discriminating crystal structures that constitute such a diverse set, and this method has been validated using both domains and protein chains.

RMSD values were used to generate dendrograms for each of the selected MSAs. The dendrograms are constructed using the hierarchical clustering algorithm built into R, `hclust` (51), with UPGMA (mean values) used as the chosen agglomeration method (52).

Each domain is assigned to its respective cluster using the assigned optimal K-values as input to Lloyd's algorithm. For each sequence group, we perform 1000 K-means clustering simulations on the MDS coordinates, and take the most common partition generated in these simulations to assign each structure to its respective cluster.

We then select a representative structure from each of the assigned clusters. The representative member for each cluster is the member with the lowest Euclidean distance to the cluster mean, using the coordinates obtained by multidimensional scaling (see description above). These cluster representatives are then taken as the distinct conformations for this protein, and are used for the binding leverage calculations and networks analyses (below).

Modeling Protein Motions by Directly Using Displacement Vectors from Alternative Conformations

As discussed, conformational changes may be modeled using vectors connecting pairs of corresponding residues in crystal structures of alternative conformations (termed "ACT", for "absolute conformational transitions"). This more direct model of conformational change is especially straightforward to apply to single-chain proteins; applying this method on a database scale to multi-chain complexes introduces confounding factors related to chain-chain correspondence between such complexes when each complex has multiple copies of a given chain.

When we use ACT to apply the modified binding leverage framework for these proteins, we observe that our surface-critical residues are significantly more conserved than are non-critical residues (Fig. S17, left), and the same trend is observed when this is applied in our dynamical network analysis for identifying interior-critical residues (Fig. S17, right). There are too few human single-chain proteins to perform a reliable analysis in which conservation is evaluated using 1000 Genomes or ExAC data – for instance, only 9 (16) structures are such that 1000 Genomes (ExAC) SNVs overlap with interior-critical residues.

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

Conservation Across Species

All cross-species conservation scores represent the ConSurf scores, as downloaded from the ConSurf Server (53-56), in which scores for each protein chain are normalized to 0. Low (i.e., negative) ConSurf scores represent a stronger degree of conservation, and high (i.e., positive) scores designate weaker conservation. We perform cross-species conservation analysis on those proteins for which ConSurf files are available from the ConSurf server, and all ConSurf scores were calculated using default parameters, listed here:

Homolog search algorithm: CSI-BLAST
Number of iterations: 3
E-value cutoff: 0.0001
Proteins database: UniRef-90
Maximum homologs to collect: 150
Maximal %ID between sequences: 95
Minimal %ID for homologs: 35
Alignment method: MAFT-L-INS-i
Calculation method: Baysian
Calculation method: JTT

Each individual point within the cross-species conservation plots (e.g., Figs. 3*B* and 3*F*, and Fig. S17) represents data from one protein: the value of the point for any given protein represents the mean conservation score for all residues within one of two classes: the set of N critical residues within a protein structure (surface or interior) or a randomly-selected set of N non-critical residues (with the same "degree", see below)

within the same structure. The randomly selected non-critical set of residues was chosen in a way such that, for each critical residue with degree k (k being the number of non-adjacent residues with which the critical residue is in contact, see below), a randomly selected non-critical residue with the same degree k was included in the set. The distributions of non-critical residues shown are very much representative of the distributions observed when re-building the random set many times.

Note that the degree (i.e., k) of residue j is defined as the number of residues which interact with residue j, where residues adjacent to residue j in sequence are not considered, and an interaction is defined whenever any heavy atom in an interacting residue is within 4.5 Angstroms of any heavy atom in residue j. We use degree as a measure of residue burial for several reasons. This metric for burial is consistent with our networks-based analysis for identifying interior-critical residues, as well as our use of residue-residue contacts in building networks for producing the ANMs. In addition, degree is also an attractive metric because it is discrete in nature, thereby allowing us to generate null distributions of non-critical residues with the exact same degree distribution.

Measures of Conservation Amongst Humans from Next-Generation Sequencing

All SNVs hitting protein-coding regions that result in amino acids changes (i.e., nonsynonymous SNVs) were collected from the phase 3 release of The 1000 Genomes Project (28). VCF files containing the annotated variants were generated using VAT (57). For nonsynonymous SNVs, the VCF files included the residue ID of the affected residue, as well as additional information (such as the corresponding allele frequency, the ancestral allele, and the residue type). To map the 1000 Genomes SNVs on to protein structures, FASTA files corresponding to the translated chain(s) of the respective transcript ID(s) were obtained using BioMart (58). FASTA files for each of the PDB structures associated with these transcript IDs (the PDB ID-transcript ID correspondence was also obtained using BioMart) were generated based on the ATOM records of the PDB files. For each given protein chain, BLAST was used to align the FASTA file obtained from BioMart with that generated from the PDB structure. The residue-residue correspondence obtained from these alignments was then used in order to map each SNV

to specific residues within the PDB. As a quality assurance mechanism, we confirmed that the residue type reported in the VCF file matched that specified in the PDB file.

ExAC SNVs were downloaded from the ExAC Browser (Beta), as hosted at the Broad Institute (29). SNVs were mapped to all PDBs following the same protocol as that used to map 1000G SNVs, and only non-synonymous SNVs in ExAC were analyzed. When evaluating SNVs from the ExAC dataset, minor allele frequencies (MAF) were used instead of DAF values. The ancestral allele is not provided in the ExAC dataset – thus, analysis is performed for MAF rather than DAF. However, we note that little difference was observed when using AF or DAF values with 1000 Genomes data, and we believe that the results with MAF values would generally be the same to those with DAF values. We also highlight the attractive feature of recapitulating the general conservation trends observed using a separate matric.

When analyzing both 1000 Genomes and ExAC data, we consider only those structures in which at least one critical and one non-critical residue are hit by a non-synonymous SNV. This enables a more direct comparison between critical and non-critical residues, as comparisons between two different proteins would rely on the assumption of equal degrees of selection between such proteins.

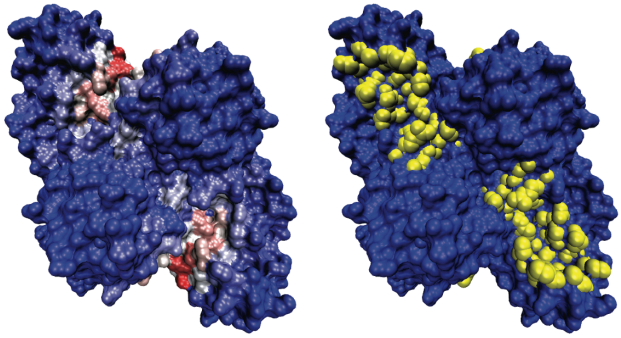
Each individual point within the intra-human conservation plots (e.g., Figs. 3*C*, 3*D*, 3*G*, and 3*H*) represents data from one protein: the value of the point for any given protein represents the mean score (DAF or MAF, for 1000 Genomes or ExAC SNVs, respectively) for all critical (red bars) or non-critical (blue bars) residues to be hit by SNVs.

The *fraction* of rare SNVs to hit a particular protein "annotation" (described below) is defined to be the ratio of the number of rare non-synonymous SNVs in that annotation to the total number of non-synonymous SNVs to hit that annotation. An annotation for a given protein is simply the set of residues within a particular category, such as the set of all surface-critical residues (or alternatively the set of all interior-critical residues, or the set of non-critical residues). We define the term "rare" to mean that a 1000 Genomes SNV has a DAF value below a certain threshold – for instance, variable thresholds ranging from DAF = 0.05% to 0.50% are evaluated in Fig. S15. An SNV in

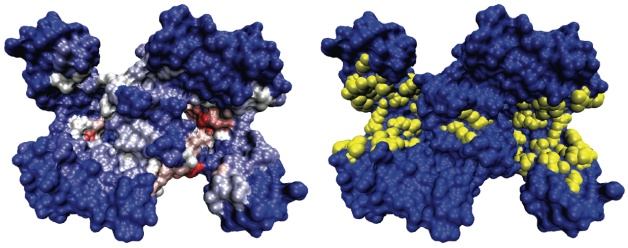
ExAC is defined to be rare if it has a MAF value below a certain threshold – variable thresholds ranging from MAF = 0.05% to 0.50% are evaluated in Fig. S16.

If a particular annotation, such as the set of surface-critical residues, has a rare SNV, then this rarity may potentially be a consequence of purifying selection acting to remove a deleterious SNV from the population pool (thereby making it rare). Such an annotation may thus be sensitive to sequence changes, and would thus be conserved. If there is a high concentration (i.e., fraction) of such rare SNVs within the annotation, it provides further confidence to the claim that the annotation is conserved. Thus, a high fraction of rare SNVs is used as a signature for stronger conservation. Supporting this intuition, previous studies have observed that conserved genomic regions within the human population harbor a higher fraction of rare SNVs (28, 30, 59).

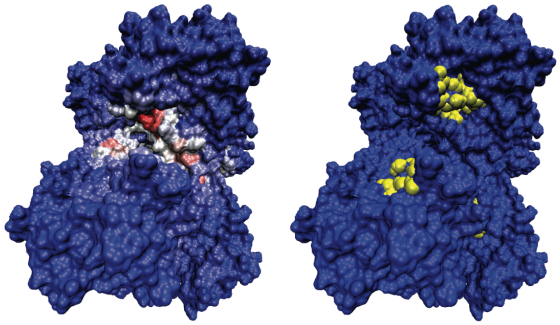
Fig. S1: Canonical proteins with surface-critical and known ligand-binding sites
Each left image shows sites that are scored highly (i.e., surface-critical residues, in red), and each right image shows the residues (yellow) that actually come into contact of known ligands, based on the corresponding *holo* structure (Supp. Table 1).



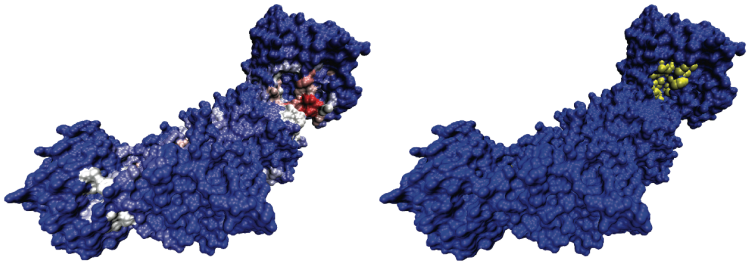
Phosphofructokinase (PDB ID 3pfk)



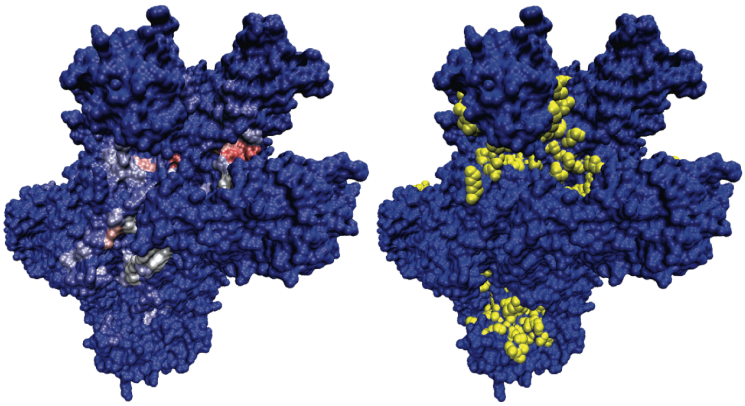
Adenylate Kinase (PDB ID 4ake)



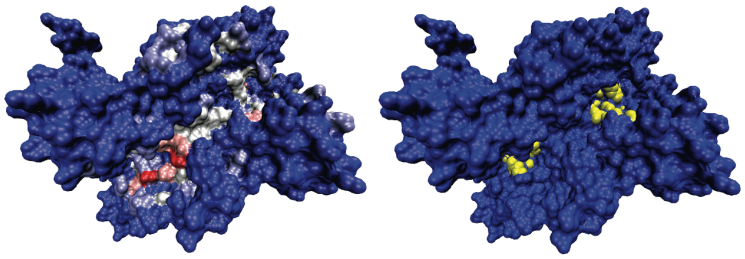
G6P-Deaminase (PDB ID 1cd5)



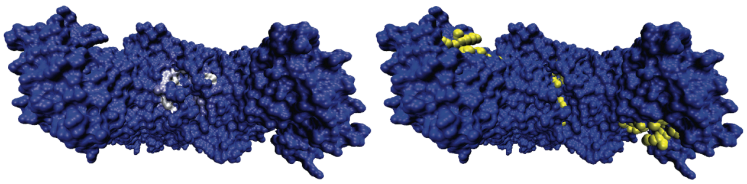
Trp Synthase (PDB ID 1bks)



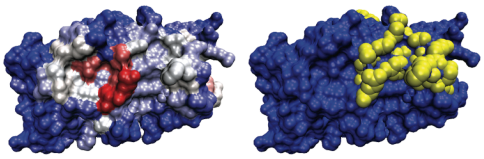
Glu Dehydrogenase (PDB ID 1nr7)



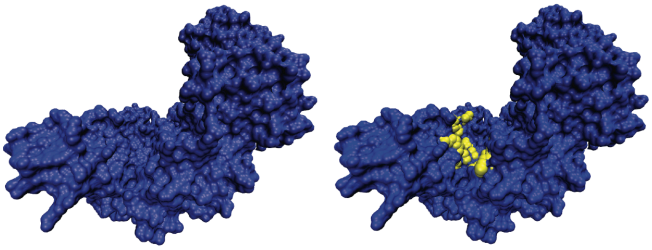
Thr Synthase (PDB ID 1e5x)



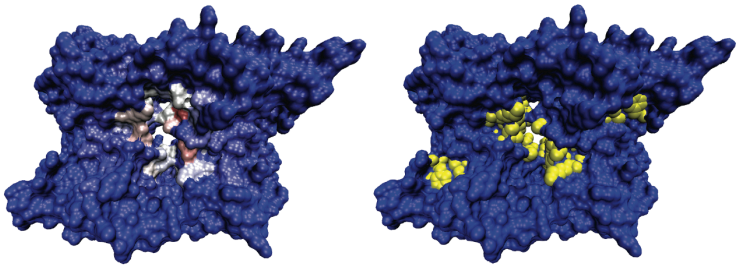
Malic Enzyme (PDB ID 1efk)



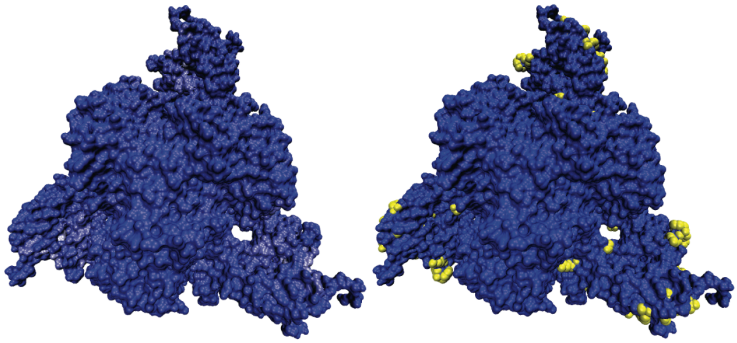
Tyr Phosphatase (PDB ID 2hnp)



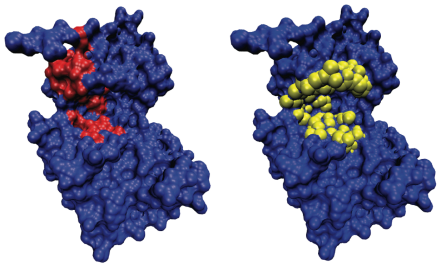
Arg Kinase (PDB ID 3ju5)



Phosphoribosyltransferase (PDB ID 1xtt)



Asp Transcarbamoylase (PDB ID 3d7s)



cAMP-dependent Kinase (PDB ID 1j3h)

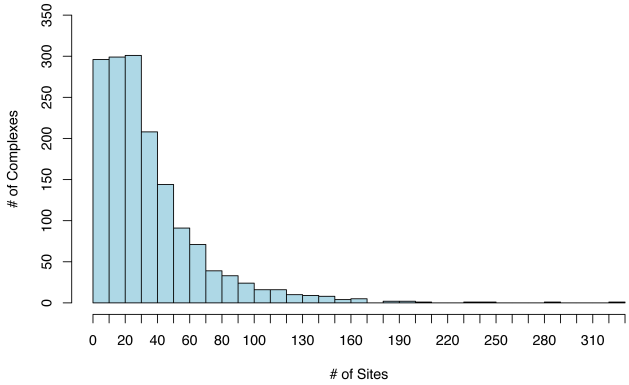


Fig. S2: Number of surface-critical sites per complex without thresholding
Complexes are taken from the the PDB biological assembly files. Shown is the distribution of the number of sites per complex. Without applying thresholds to the list of ranked surface-critical sites, the protein is often covered with an excess of identified critical sites.

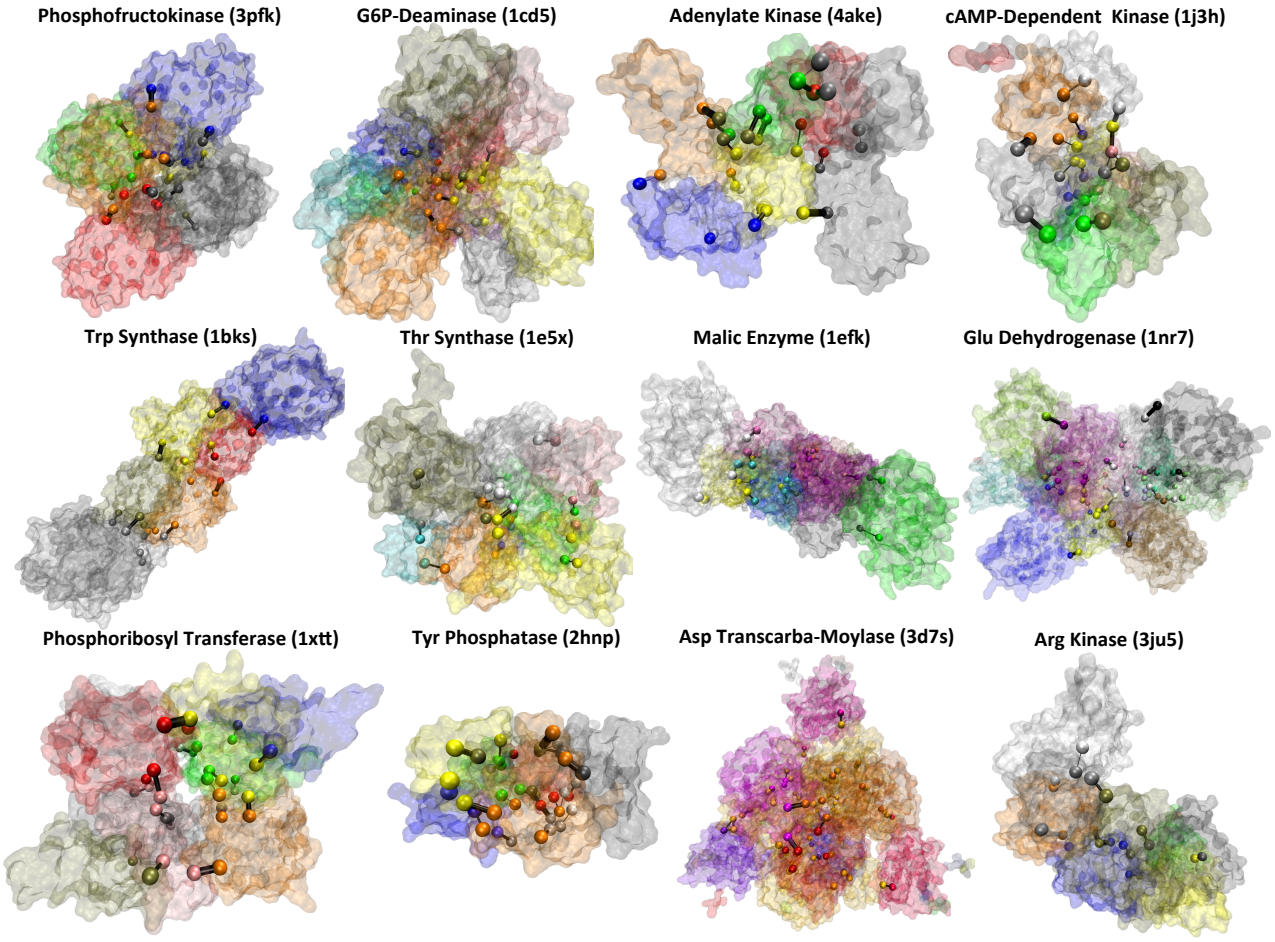


Fig. S3: Community partitioning for canonical systems
Different network communities are colored differently, and communities were identified using the dynamical network-based analysis with the GN formalism discussed in the main text and SI Methods. Residues shown as spheres are interior-critical residues, and they are colored based on community membership, and black lines connecting pairs of critical residues represent the highest-betweenness edges between the corresponding communities.

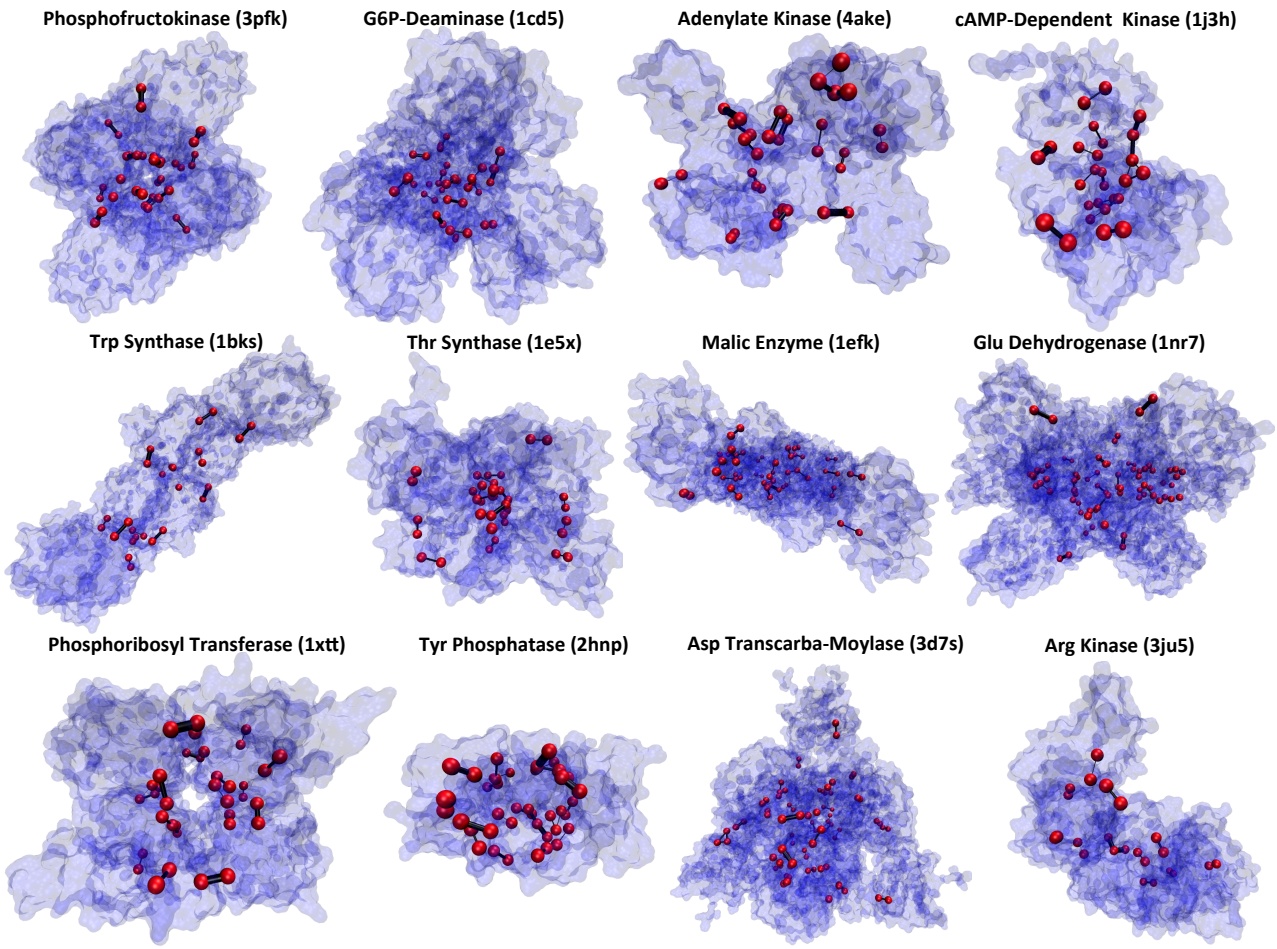


Fig. S4: Interior-critical residues highlighted in canonical systems
Shown above are the same proteins shown in Fig. S3, but with interior-critical residues highlighted in red spheres.

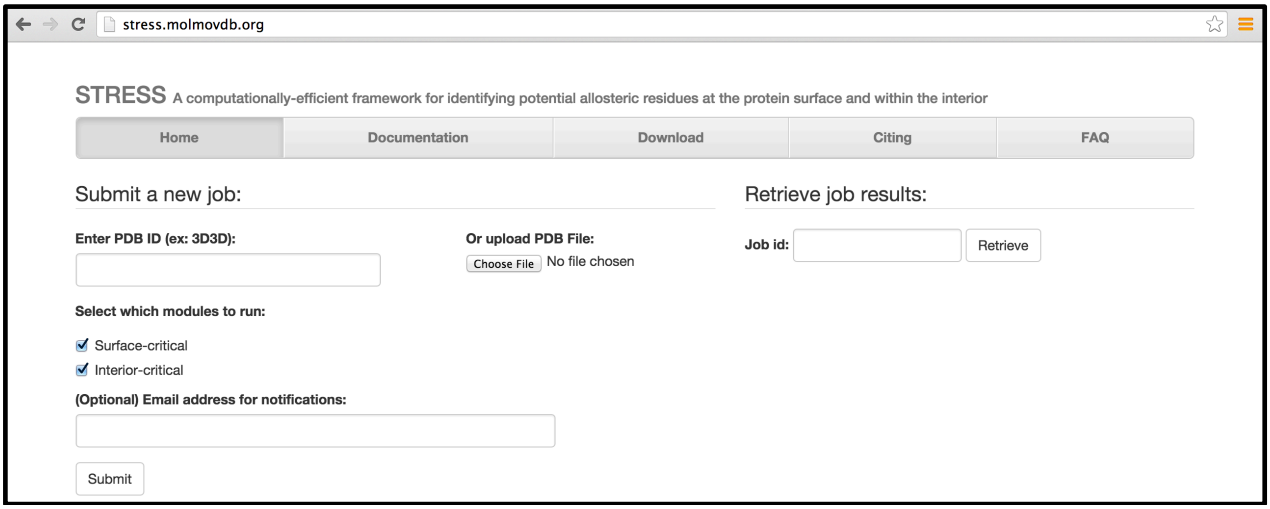


Fig. S5: Home page of the STRESS server (stress.molmovdb.org)
The server enables users to either provide PDB IDs or to upload their own PDB files for proteins of interest. Users may opt to identify surface-critical residues, interior-critical residues, or both.

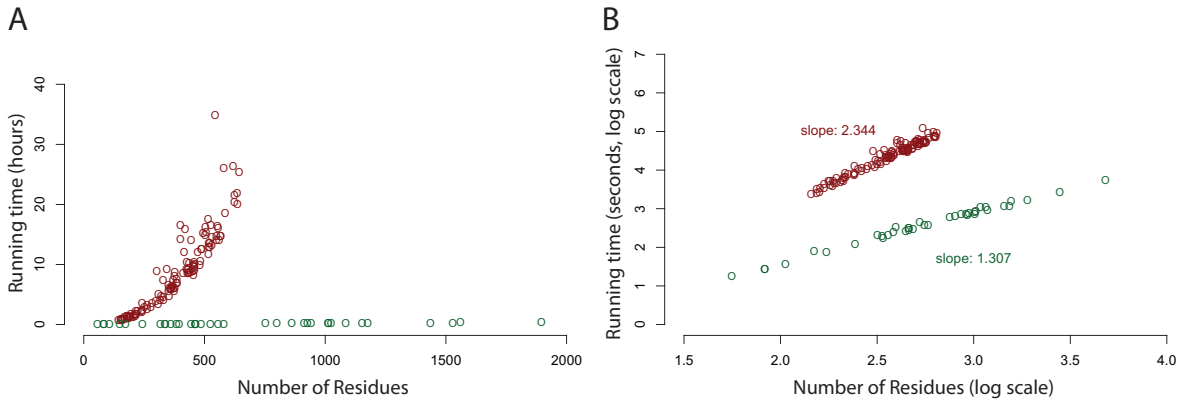


Fig. S6: Running times and optimization in the search of surface-critical residues
(A) Running times are shown for systems of various sizes. Shown in red are the running times without optimizing for speed. Performing local searching supported with hashing and implementing additional algorithmic optimizations for computational efficiency reduce running times considerably (in green), relative to a more naïve approach without optimization (in red). *(B)* The same data is represented as a log-log plot. The slopes of these two approaches demonstrate that our algorithm reduces the computational complexity by an order of magnitude. Our speed-optimized algorithm scales at $O(n^{1.3})$, where n is the number of residues.

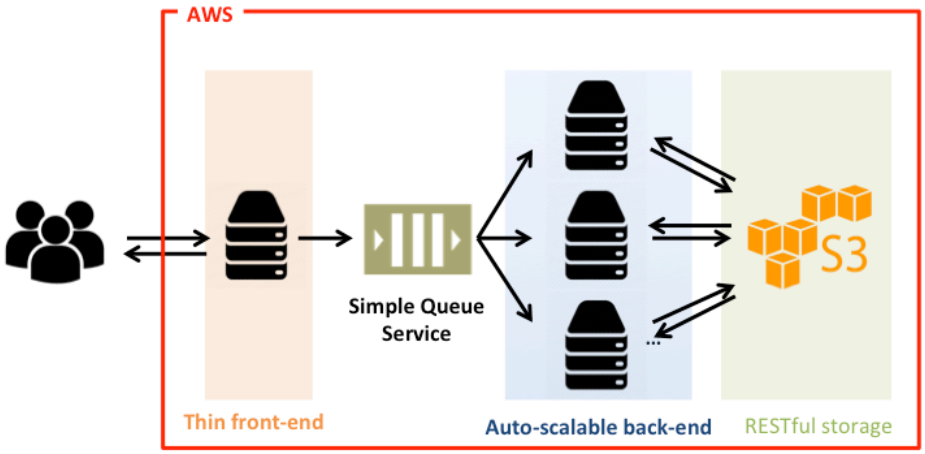


Fig. S7: Architecture of the STRESS server
A thin front-end server handles incoming user requests, and more powerful back-end servers perform the heavier algorithmic calculations. The back-end servers are dynamically scalable, making them capable of handling wide fluctuations in user demand. Amazon's Simple Queue Service is used to coordinate between user requests at the front end and the back-end compute nodes: when the front-end server receives a request, it adds the job to the queue, and back-end servers pull that job from the queue when ready. Source code is available through Github (github.com/gersteinlab/STRESS).

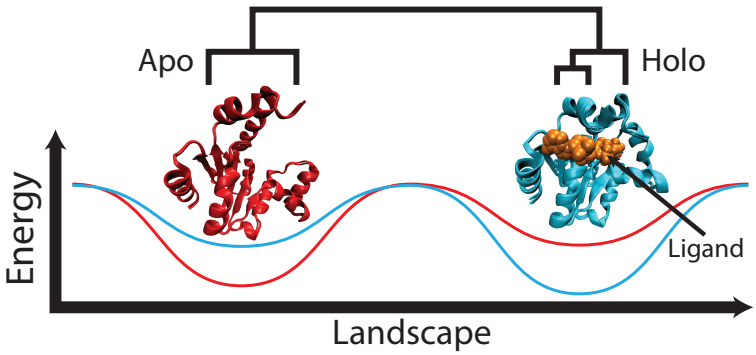


Fig. S8: Energy landscapes to describe distributions of different conformations
Energy landscape theory may be used to describe the relative populations of alternative biological states and conformations (for instance, active/inactive, or holo/apo). In the apo state, the landscape may take the form of the red curve, resulting in most proteins favoring the conformation shown in red. Once binding to ligand, the landscape becomes reconfigured to take the shape in the cyan curve, thereby shifting the distribution of conformations to that shown in cyan. One may use multiple structure alignments for domains or proteins to identify these distinct biological states in a database of structures. The schematized dendrogram represents the partitioning of these structures by a metric such as RMSD. The example shown is a multiple structure alignment of adenylate kinase. SCOP IDs of the *apo* domains: d4akea1 and d4akeb1; those of the *holo* domains: d3hpqb1, d3hpqa1, d2eckb1, d2ecka1, d1akeb1, and d1akea1.

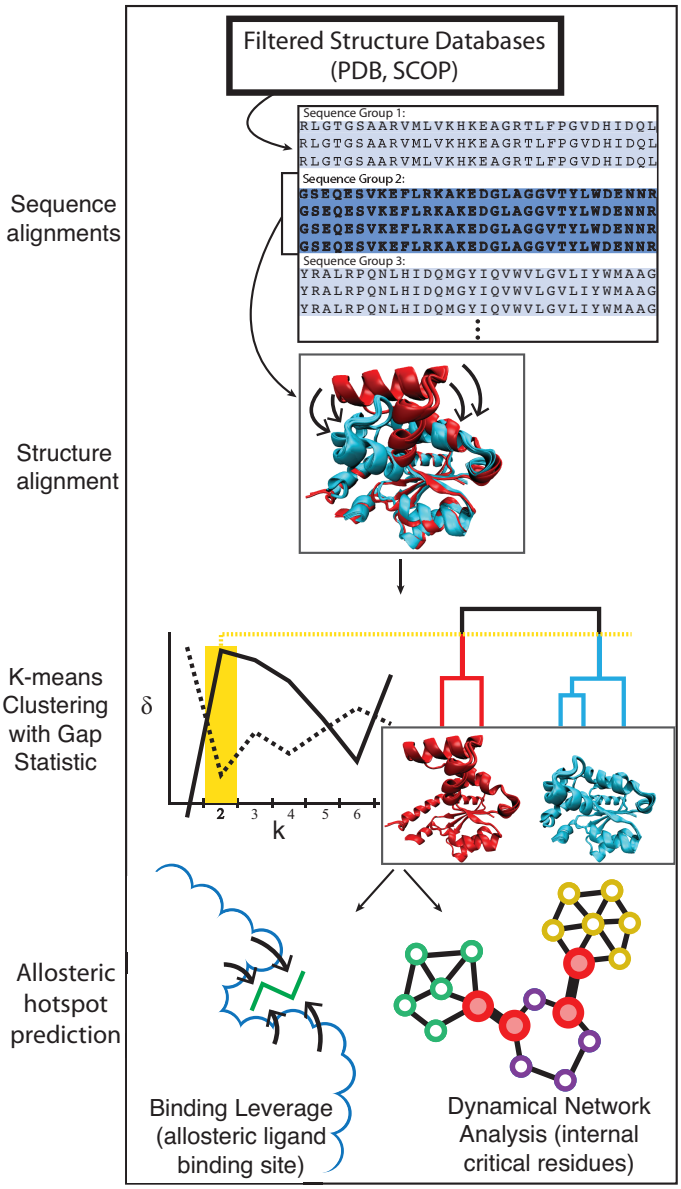


Fig. S9: Pipeline for identifying distinct conformations and critical residues
Top to bottom: BLASTClust is applied to the sequences corresponding to a filtered set of structures, thereby providing a large number of sequence-identical sets of proteins (i.e., "sequence groups"). For each sequence-identical group, a multiple structure alignment is performed using STAMP. The example shown here is adenylate kinase; details are provided in Fig. S8. Using the pairwise RMSD values in this structure alignment, the structures are clustered using the UPGMA algorithm, and K-means with the gap statistic (δ) is performed to identify the number of distinct conformations. The plot at left identifies 2 as the optimal value for K: the solid line represents δ(K) values at each value of K, and the dotted line represents $\delta(K+1) - s_{k+1}$ for each value of K (see SI Methods for details). The structures that exhibit multiple clusters (i.e., those with K > 1) are then taken to exhibit multiple conformations. Finally, surface-critical (bottom-left) and interior-critical (bottom-right) residues are identified on those proteins determined to exist as multiple conformations.

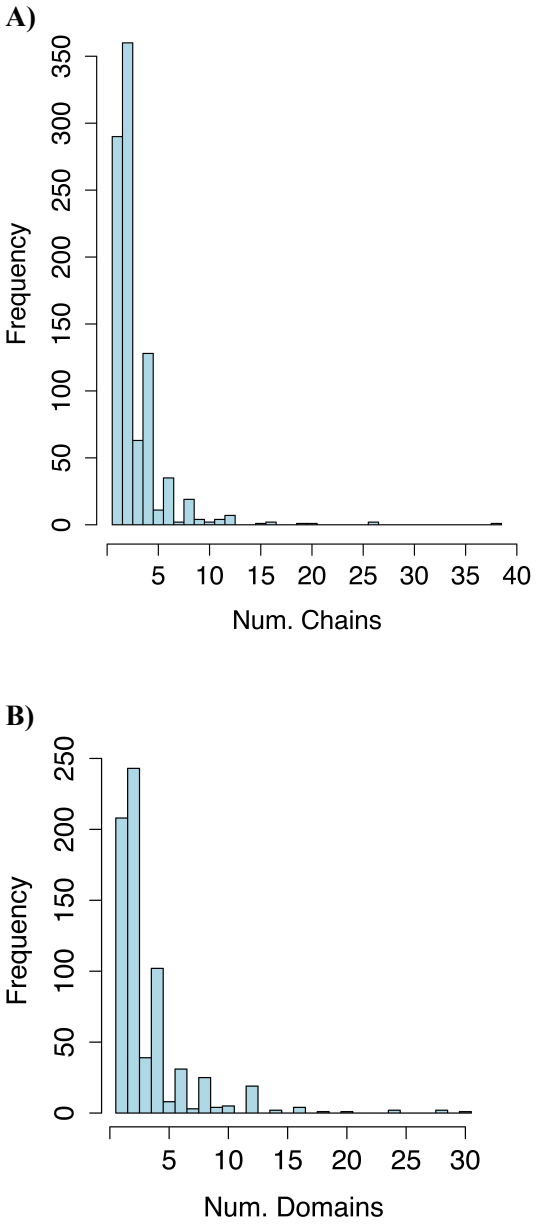
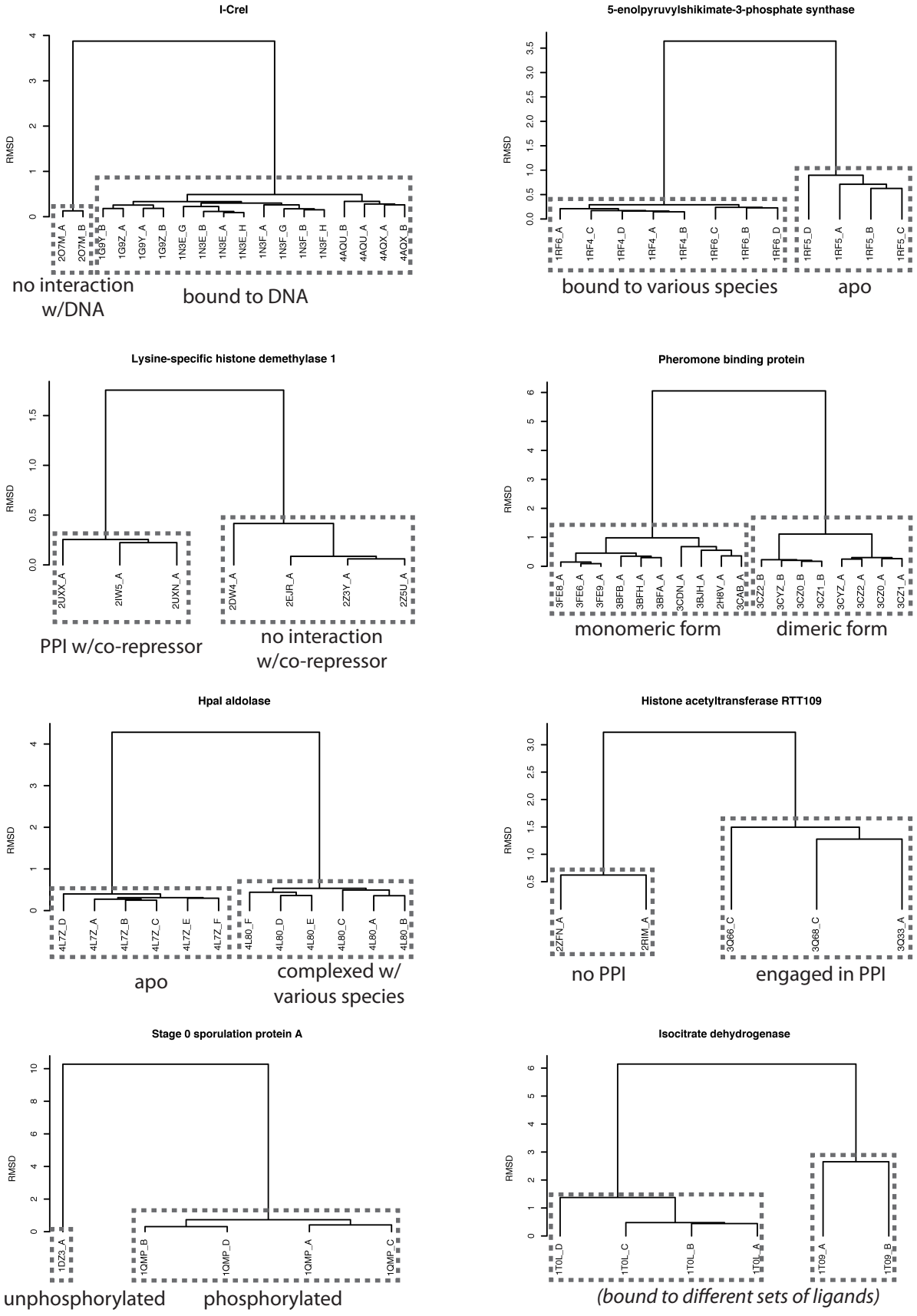
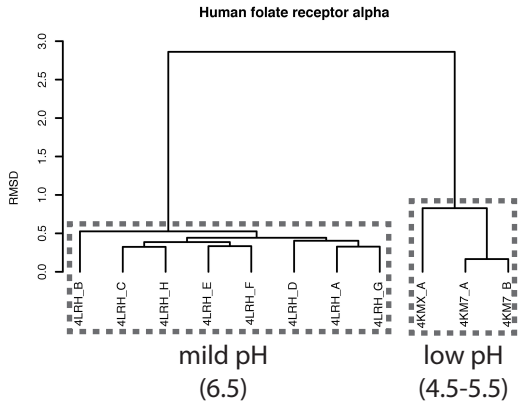


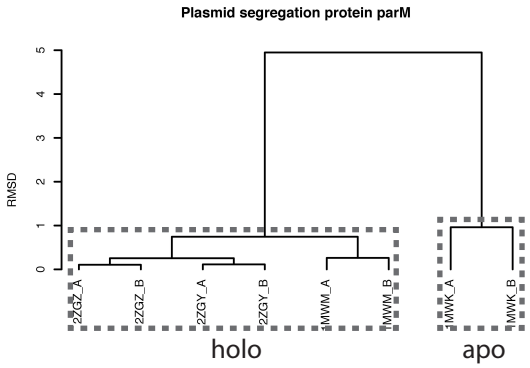
Fig. S10: Distributions of the number of chains and domains in the dataset of alternative conformations

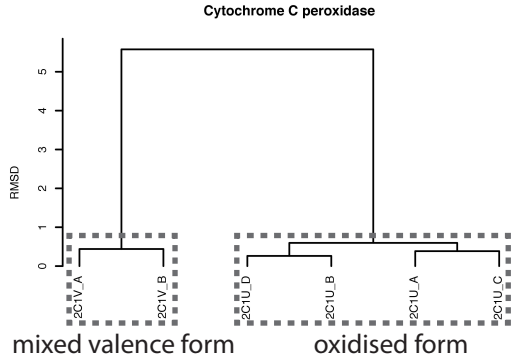
Each structure the dataset of alternative conformations is taken from the first biological assembly file of the corresponding PDB. The structures in this database vary considerably in terms of size. Shown in panel *(A)* is the histogram representing the distribution for the number of chains in these biological assemblies, and shown in *(B)* is the corresponding distribution for the number of SCOP domains.

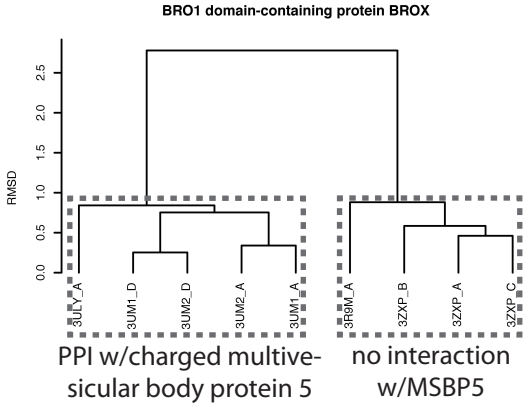
Supp. Fig. 11: Capturing alternative conformations in diverse biological contexts

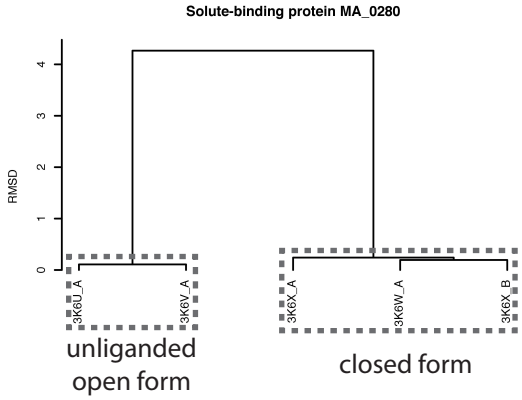


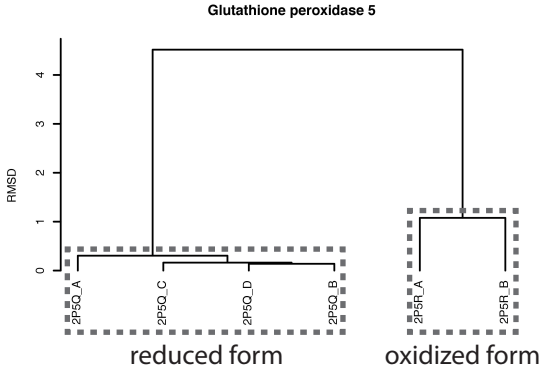

Human folate receptor alpha
RMSD
4LRH_B
4LRH_C
4LRH_H
4LRH_E
4LRH_F
4LRH_D
4LRH_A
4LRH_G
4KMX_A
4KM7_A
4KM7_B
mild pH
(6.5)
low pH
(4.5-5.5)

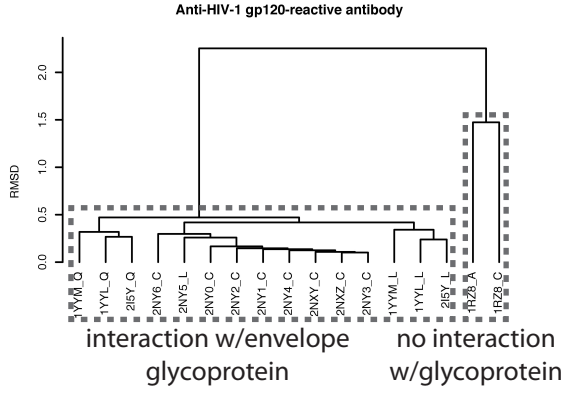

Plasmid segregation protein parM
RMSD
2ZGZ_A
2ZGZ_B
2ZGY_A
2ZGY_B
MWM_A
MWM_B
MWK_A
MWK_B
holo
apo

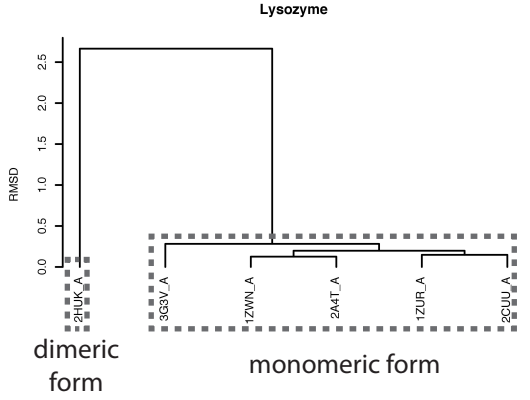

Cytochrome C peroxidase
RMSD
2C1V_A
2C1V_B
2C1U_D
2C1U_B
2C1U_A
2C1U_C
mixed valence form
oxidised form


BRO1 domain-containing protein BROX
RMSD
3ULV_A
3UM1_D
3UM2_D
3UM2_A
3UM1_A
3RBM_A
3ZXP_B
3ZXP_A
3ZXP_C
PPI w/charged multive-
sicular body protein 5
no interaction
w/MSBP5


Solute-binding protein MA_0280
RMSD
3K6U_A
3K6V_A
3K6X_A
3K6W_A
3K6X_B
unliganded
open form
closed form


Glutathione peroxidase 5
RMSD
2P5Q_A
2P5Q_C
2P5Q_D
2P5Q_B
2P5R_A
2P5R_B
reduced form
oxidized form


Anti-HIV-1 gp120-reactive antibody
RMSD
1YYM_C
1YYL_C
2I5Y_C
2NY6_C
2NY5_L
2NY0_C
2NY2_C
2NY1_C
2NY4_C
2NXY_C
2NXZ_C
2NY3_C
1YYM_L
1YYL_L
2I5Y_L
1HZ8_A
1HZ8_C
interaction w/envelope
glycoprotein
no interaction
w/glycoprotein


Lysozyme
RMSD
2HUK_A
3G3V_A
1ZWN_A
2A4T_A
1ZUR_A
2CUU_A
dimeric
form
monomeric form

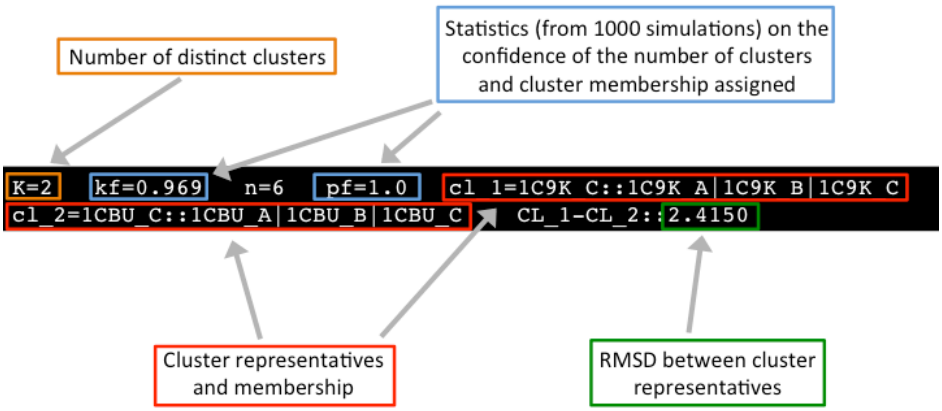
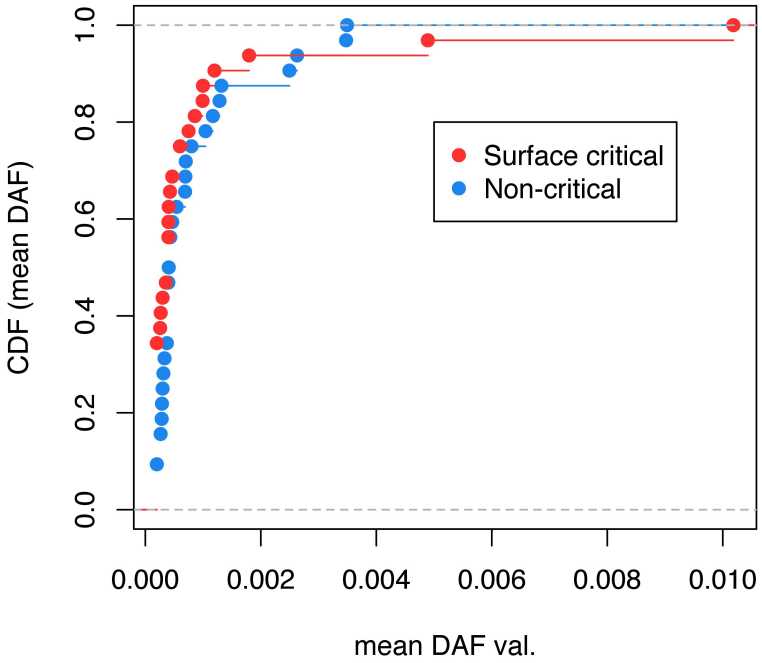
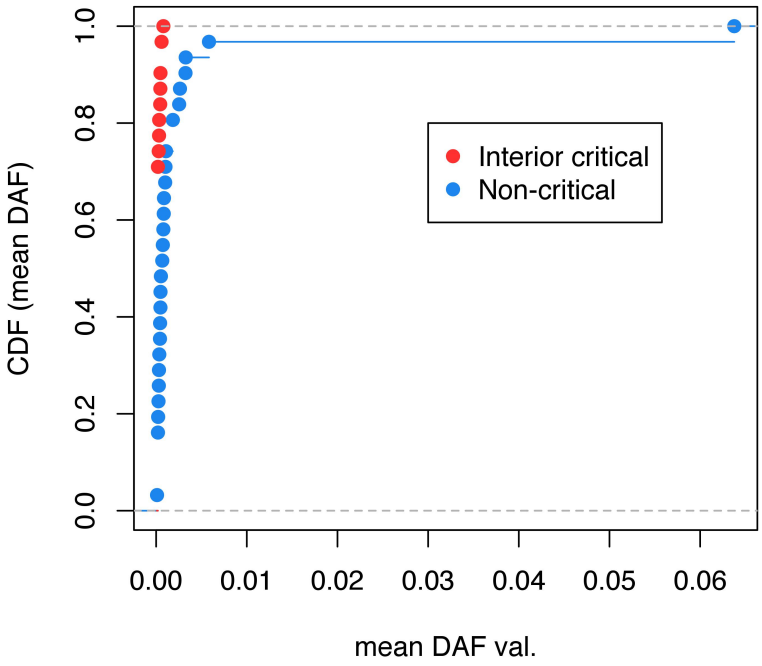


Fig. S12: A single annotated entry from our database of alternative conformations
The clustering for the protein adenosylcobinamide kinase is shown. Two distinct conformations are represented in the ensemble of structures. The measure *kf* designates the fraction of times that the optimal value of K (here, K=2) was obtained out of 1000 simulations in which the algorithm (K-means with the gap statistic) obtained this particular value of K. The high *kf* value (0.969) signifies that these structures are very well clustered into two groups. *n* designates the number of distinct structures (PDB chains in this case) in the multiple structure alignment. *pf* designates the fraction of times (out of 1000 simulations of running Lloyd's algorithm, the standard K-means algorithm) that this particular set of structure-group assignments were assigned. In this this example, for all 1000 simulations, 1C9K_C and 1C9K_A were clustered in one group, and 1CBU_A, 1CBU_B, 1CBU_C clustered together. Within each cluster (the two clusters shown as two red boxes), the chain preceding the "::" tag designates the cluster representative (i.e., the structure closest to the Euclidean centroid of the cluster). The last field gives the RMSD values between cluster representatives. See the header information within Supp. File 1 for further details.

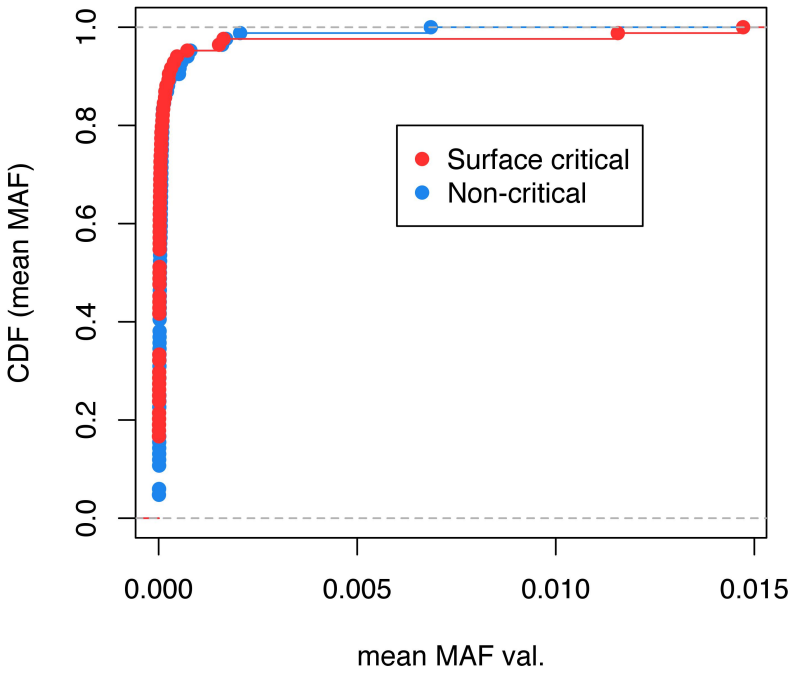


A) Cumulative distribution functions for mean DAF values of surface-critical and non-critical residues (p-val = 0.159, KS test)

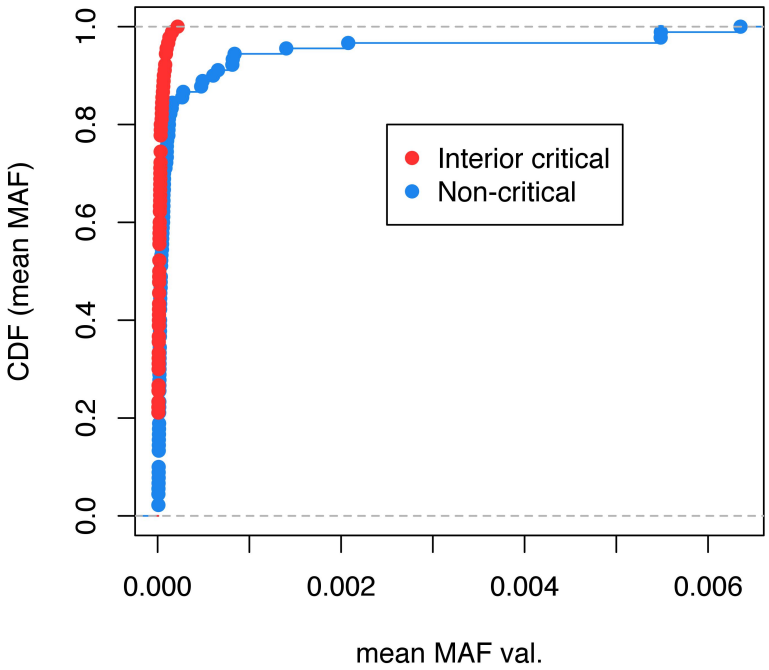


B) Cumulative distribution functions for mean DAF values of interior-critical and non-critical residues (p-val = 1.79e-4, KS test)

Fig. S13: Potential shifts in DAF distributions (in 1000 Genomes) using two-sample Kolmogorov-Smirnov tests



A) Cumulative distribution functions for mean minor allele frequencies of surface-critical and non-critical residues (p-val = 9.49e-2, KS test)



B) Cumulative distribution functions for mean minor allele frequencies of interior-critical and non-critical residues (p-val = 1.75e-4, KS test)

Fig. S14: Potential shifts in mean minor allele frequency distributions (in ExAC) using two-sample Kolmogorov-Smirnov tests

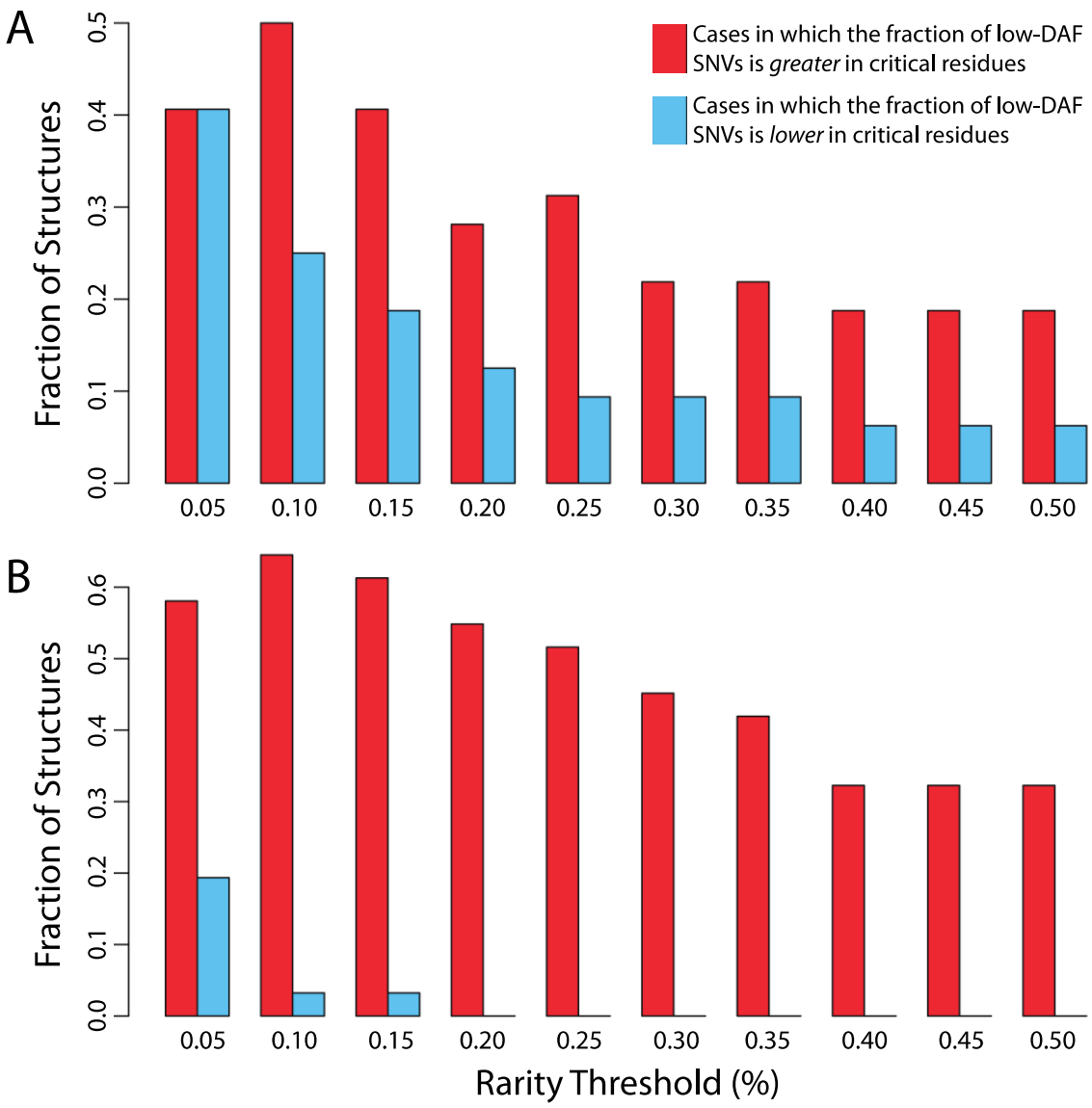


Fig. S15: Measuring relative conservation by the fraction of rare (low-DAF) variants using 1000 Genomes data

Protein regions with high fractions of *rare* variants are believed to be more sensitive to sequence variants than are other regions (thereby explaining why such variants occur infrequently in the population). Here, a rarely occurring SNV within the human population is defined to be one with a DAF less than or equal to the rarity thresholds given on the y-axis. We consider all structures such that at least one critical and at least one non-critical residue are hit by a 1000 Genomes non-synonymous SNV. Distributions in which the critical residues are defined to be the surface-critical *(A)* and interior-critical *(B)* residues are shown. For varying thresholds to define rarity, there are more structures in which the fraction of rare variants is higher in critical residues than in non-critical residues. Cases in which the fraction is equal in both categories are not shown. *(A)* represents data from 31 structures, and *(B)* represents data from 32 structures.

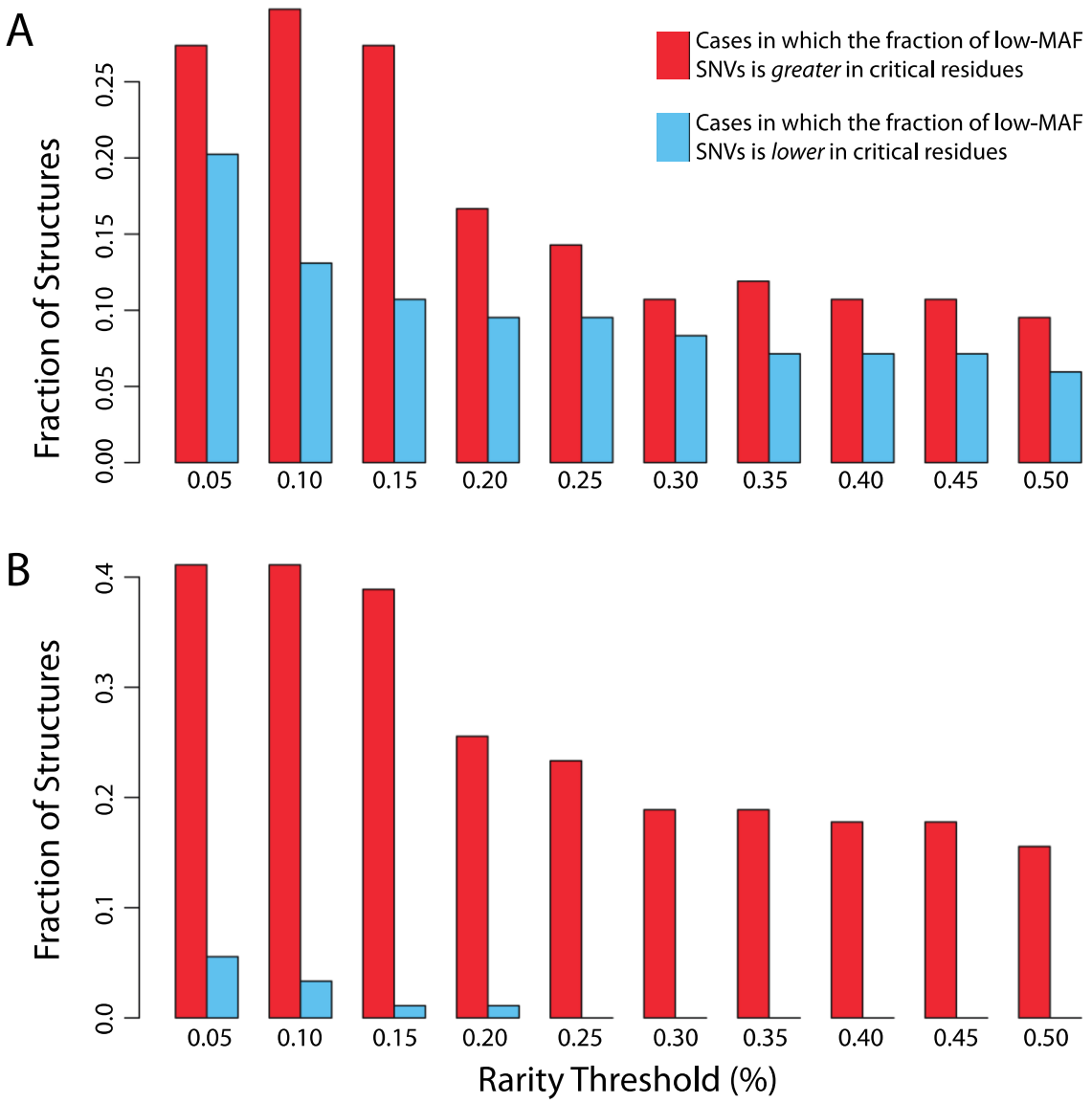


Fig. S16: Measuring relative conservation by the fraction of rare (low-MAF) variants using ExAC data

Protein regions with high fractions of *rare* variants are believed to be more sensitive to sequence variants than are other regions (thereby explaining why such variants occur infrequently in the population). Here, a rarely occurring SNV within the human population is defined to be one with a MAF less than or equal to the rarity thresholds given on the y-axis. We consider all structures such that at least one critical and at least one non-critical residue are hit by a non-synonymous SNV in the ExAC dataset. Distributions in which the critical residues are defined to be the surface-critical *(A)* and interior-critical *(B)* residues are shown. For varying thresholds to define rarity, there are more structures in which the fraction of rare variants is higher in critical residues than in non-critical residues. Cases in which the fraction is equal in both categories are not shown. *(A)* represents data from 90 structures, and *(B)* represents data from 84 structures.

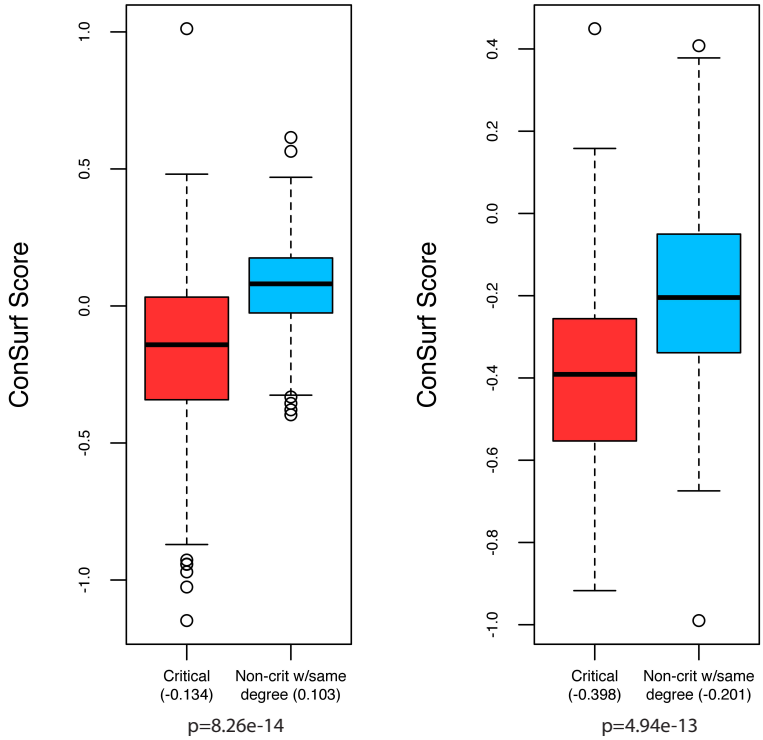


Fig. S17: Modeling protein conformational change through a direct use of crystal structures from alternative conformations using absolute conformational transitions (ACT)
Left: Distributions (155 structures) of the mean conservation scores on surface-critical (red) and non-critical residues with the same degree of burial (blue). *Right*: Distributions (159 structures) of the mean conservation scores for interior-critical (red) and non-critical residues with the same degree of burial (blue). Mean values are given in parentheses. Results for single-chain proteins are shown, and p-values were calculated using a Wilcoxon rank sum test.

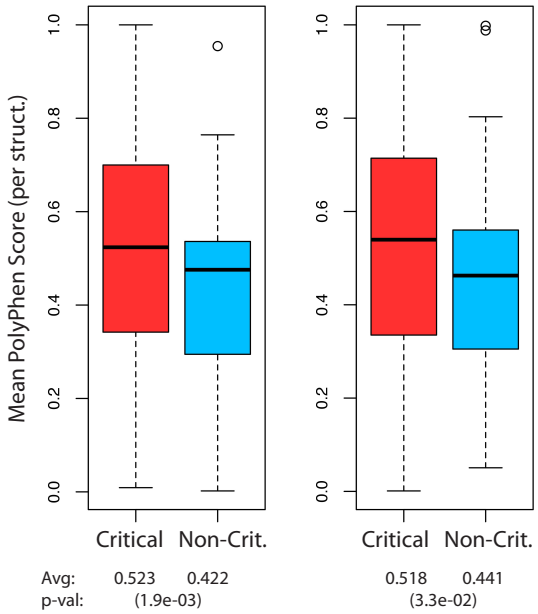


Fig. S18: Mean PolyPhen scores for critical- and non-critical residues, as identified by ExAC

Left: Distributions (64 structures) of mean PolyPhen values on surface-critical residues (red) and non-critical residues (blue). *Right*: Distributions (70 structures) of mean PolyPhen values on interior-critical residues (red) and non-critical residues (blue). Overall mean values and p-values are given below plots. Note that higher PolyPhen scores denote more damaging variants.

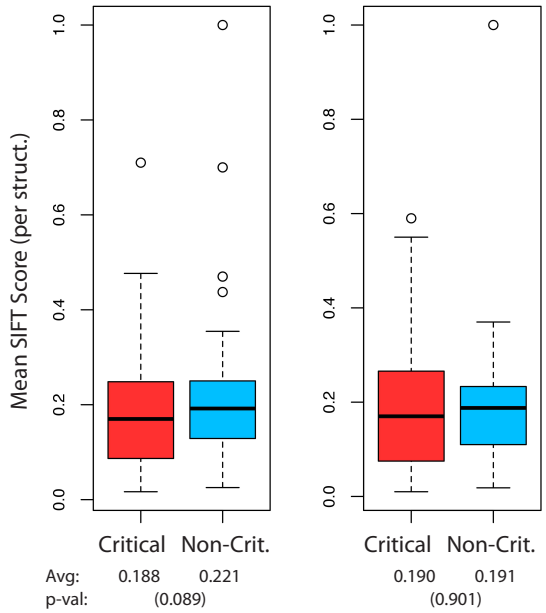


Fig. S19: Mean SIFT scores for critical- and non-critical residues, as identified by ExAC

Left: Distributions (63 structures) of mean SIFT values on surface-critical residues (red) and non-critical residues (blue). *Right*: Distributions (65 structures) of mean SIFT values on interior-critical residues (red) and non-critical residues (blue). Overall mean values and p-values are given below plots. Note that lower SIFT scores denote more damaging variants.

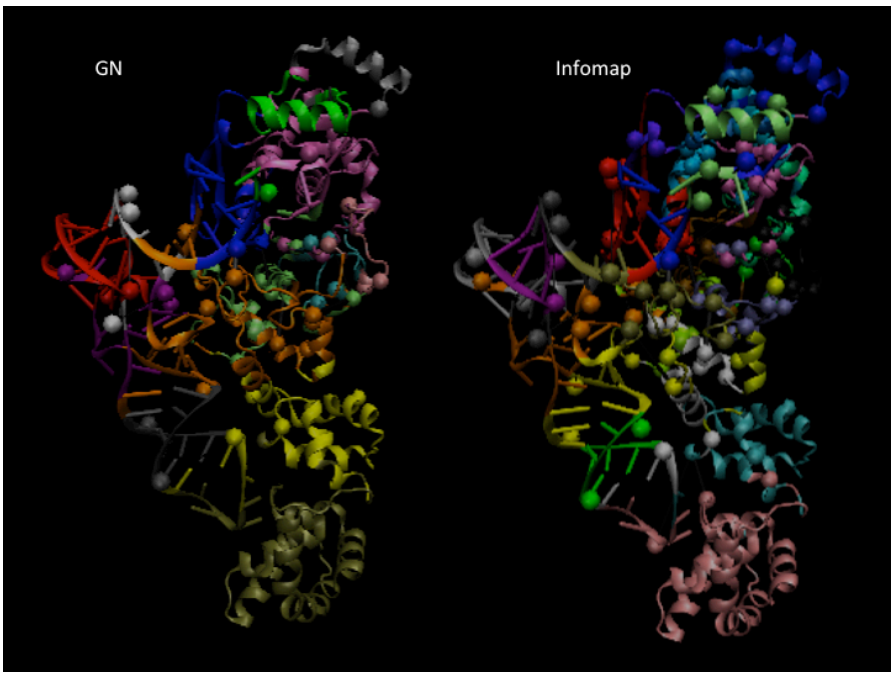


Fig. S20: Network modularization by GN and Infomap
Different colors correspond to different communities. Network modularization by the GN (left) and Infomap (right) algorithms are shown for the crystal structure of glutamyl-tRNA synthetase complexed with tRNA(Glu) and glutamol-AMP (PDB 1N78).

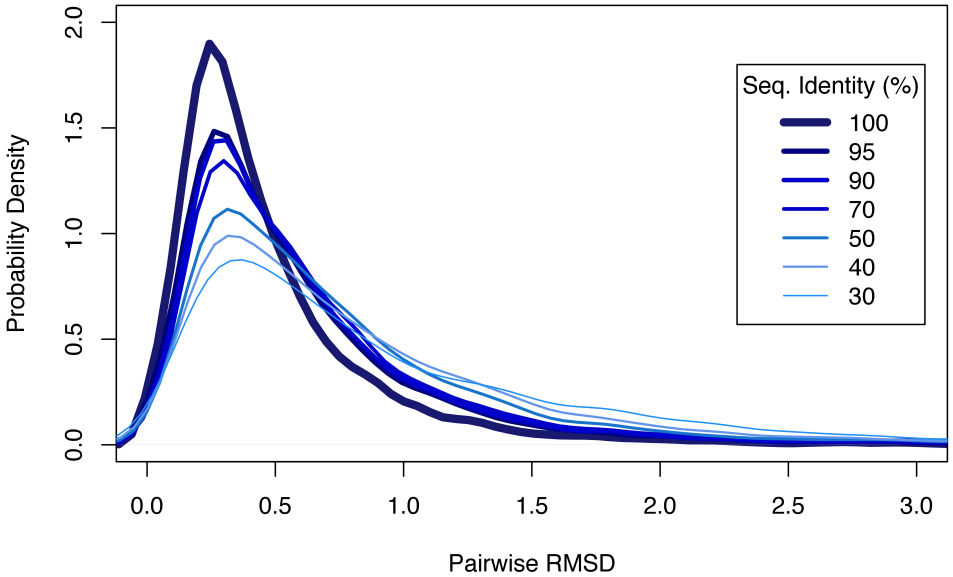
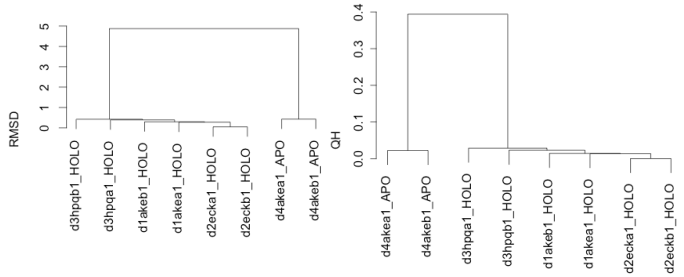


Fig. S21: Probability distributions of pairwise-RMSD by sequence identity
Distributions for average pairwise RMSD values across domains within all multiple structure alignments at varying levels of sequence identity.

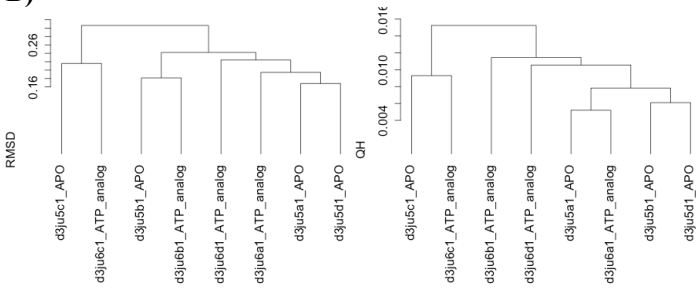
Fig. S22: **Representative clustering of domains based on RMSD and Q_H: RMSD generally matches the clustering obtained when using Q_H**

Shown are the dendrograms for domains in adenylate kinase *(A)*, arginine kinase *(B)*, calcyclin *(C)*, and catabolite activator protein *(D)*

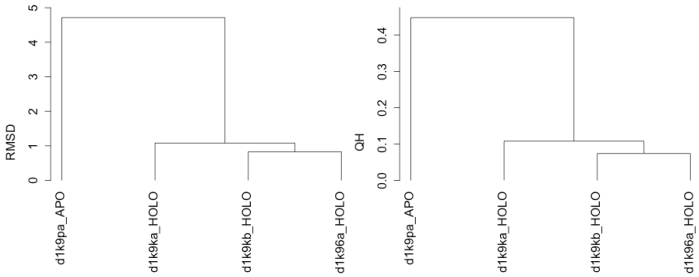
A)



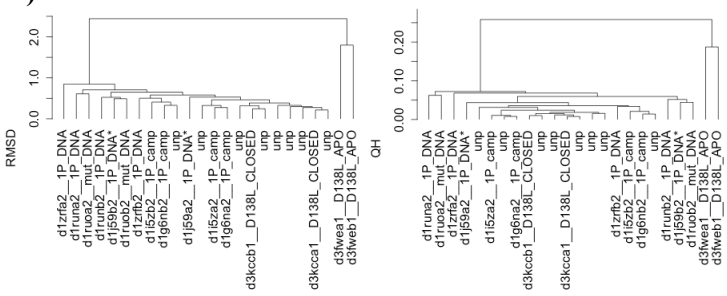
B)



C)



D)



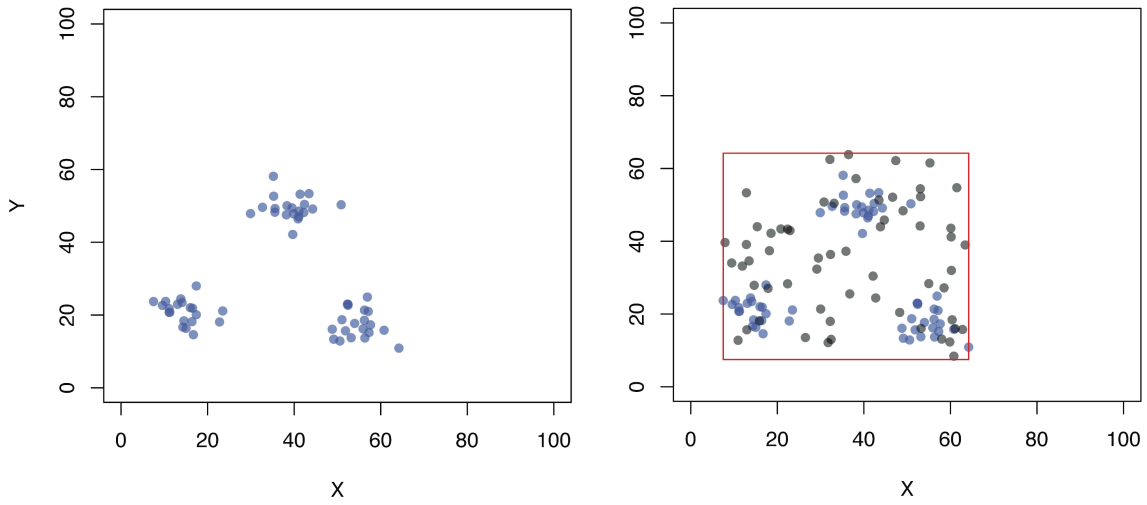


Fig. S23: Intuition behind the k-means algorithm with the gap statistic
The objective is to identify the ideal number of clusters to describe the observed data of 60 points (in blue). This entails defining how well-clustered our observed data appears (given an assigned number of clusters, K) relative to a null model consisting of a randomly distributed set of 60 points (grey) that fall within the same variable ranges as the observed data. Further details are provided by Tibshirani et al, 2001.

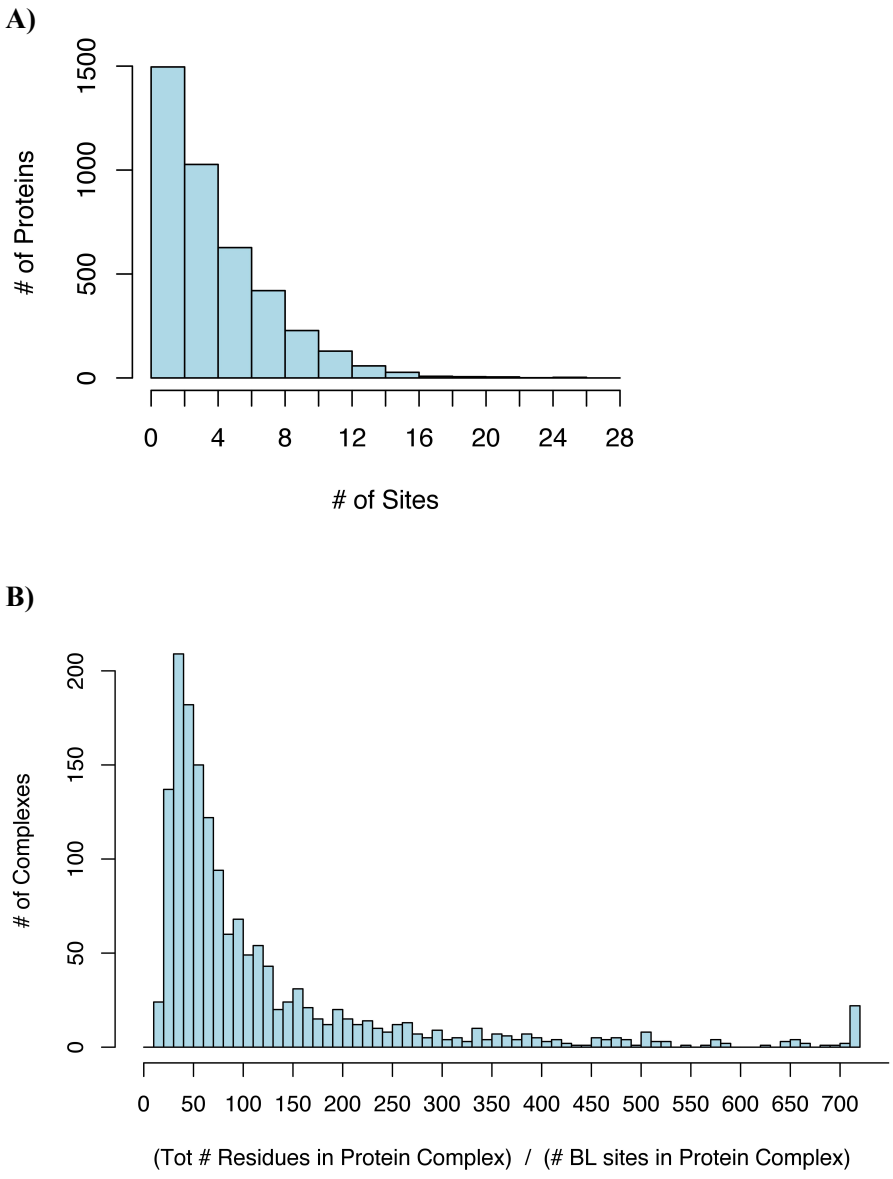


Fig. S24: Quantifying the number of distinct surface-critical sites
(A) The distribution of the number of surface-critical sites per PDB chain; *(B)* The density of surface-critical residues with respect to the total number of residues in the biological assembly (here referred to as a "complex", though in some cases, the biological assembly may in fact be a single chain).

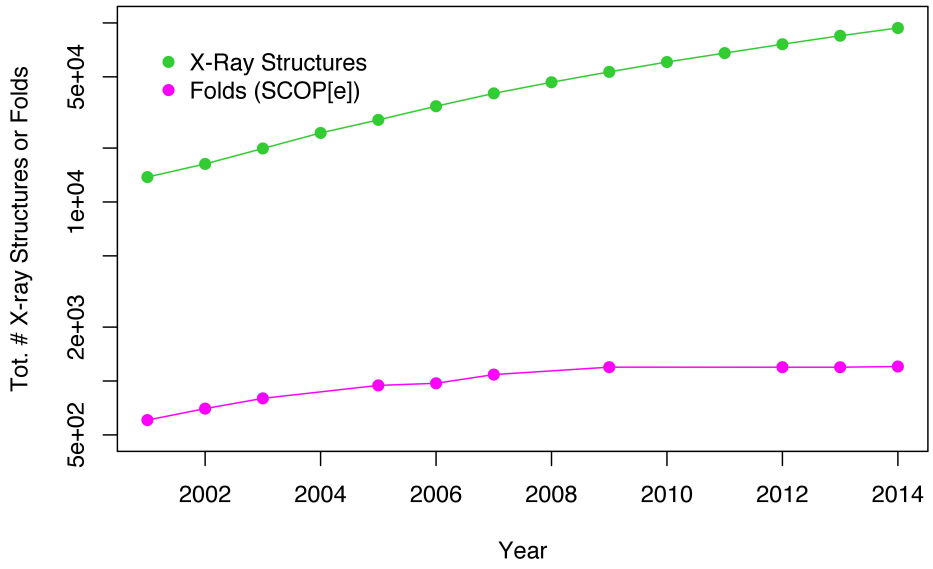


Fig. S25: Growth rate of deposited PDB structures, and the concomitant growth rate in the number of folds (as defined by CATH and SCOP)
The growing appreciation for dynamic behavior and the importance of conformational heterogeneity is being facilitated by a growing redundancy within the PDB. Such redundancy is represented, for instance, when the same protein is structurally resolved under different conditions, potentially resulting in alternative conformations.

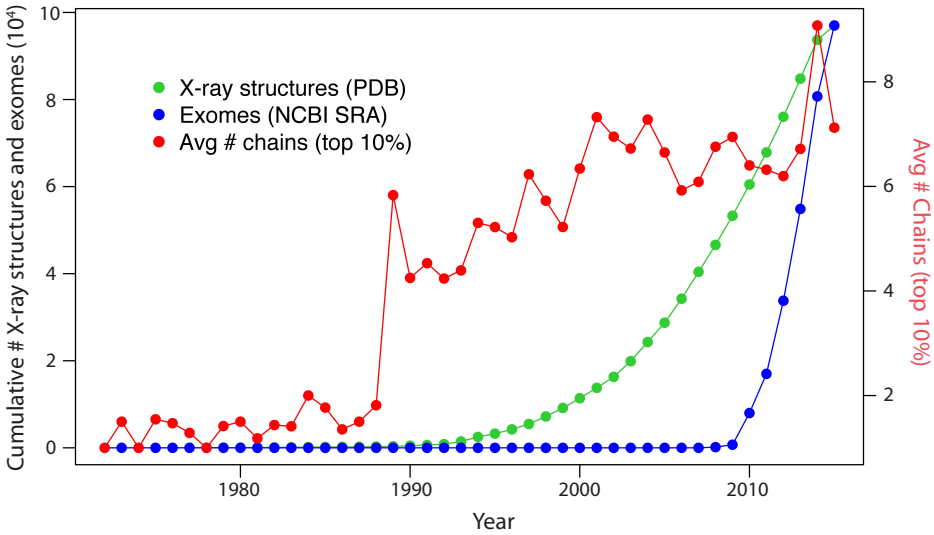


Fig. S26: Trends in data generation point to growing opportunities for leveraging sequence variants to study structure (and vice versa)
The volume of sequenced exomes is outpacing that of structures, while solved structures have become more complex in nature. Red: Average number of chains per PDB (considering the biological assembly PDB files for the top 10% of PDBs for a each year). Green: Cumulative number of X-Ray structures deposited in the PDB. Blue: Cumulative number of exomes stored in the NCBI Sequence Read Archive (SRA). All data were downloaded in May 2015.

HOLO	APO
1ake (**AP5**)	4ake
3cep (**G3P, IDM, PLP**)	1bks (**PLP**)
1hor (**AGP**, *PO4*, [& **16G** in pdb 1HOT])	1cd5
2c2b (**SAM**, [& **LLP** in pdb 2c2g])	1e5x
1gz3 (**ATP, FUM**, *OXL*)	1efk (MAK)
1atp (**ATP**)	1j3h
1hwz (**GLU**, **GTP, NDP** [& **ADP** in PDB 1NQT])	1nr7
1xtu (**CTP, U5P**)	1xtt (*ACY*, **U5P**)
1aax **(BPM** [& **892** in PDB 1T49])	2hnp
7at1 (**ATP**, **MAL**, **PCT** [& **CTP** in PDB 1RAC], [& **PAL** in PDB 1D09])	3d7s
3ju6 (**ANP**, **ARG**)	3ju5
6pfk (PGA [& **F6P + ADP** in PDB 4PFK])	3pfk (*PO4*)

Table S1: Set of 12 canonical proteins, organized by state (*apo* or *holo*)
Ligands are given in parentheses (those in bold text designate the ligand used to define residues involved in canonical ligand-binding interactions).

PDB	Fract Protein Hit	Fract Known Bio Sites Hit
3pfk	0.108	1
4ake	0.287	1
1cd5	0.09	0.5
1j3h	0.06	1
1bks *	0.132	0.25
1e5x	0.151	0.667
1efk	0.032	0
1nr7	0.071	0.75
1xtt	0.111	1
2hnp	0.672	1
3d7s *	0.052	0
3ju5	0.017	0
Avgs	**0.15 (0.16 w/o 3d7s)**	**0.60 (0.65 w/o 3d7s)**

Table S2: Identifying known ligand-binding sites
The 2nd column designates the fraction of residues that constitute surface-critical residues, and the 3rd column represents, for each structure, the fraction of known ligand-binding sites that strongly overlap with surface-critical sites.

PDB	Fract protein hit by our predictions	Fract protein actually occupied by biological ligand-binding sites
3pfk	0.11	0.13
4ake	0.29	0.17
1cd5	0.09	0.08
1j3h	0.06	0.08
1bks	0.13	0.07
1e5x	0.15	0.09
1efk	0.03	0.08
1nr7	0.07	0.16
1xtt	0.11	0.16
2hnp	0.67	0.11
3d7s *	0.05	0.07
3ju5	0.02	0.08

Table S3: Do surface-critical sites occupy an exceedingly large fraction of the protein?
For most proteins in the canonical set, the fraction of the protein occupied by surface-critical residues roughly matches the fraction of residues known to be directly involved in ligand binding. For most proteins (blue), the fraction of critical-surface residue is actually lower than that of known ligand-binding residues.

n	Fract Known Bio Sites Hit (w/ and w/o 3d7s)
6	0.60 (0.65)
5	0.62 (0.67)
4	0.69 (0.75)
3	0.74 (0.76)
2	0.81 (0.81)
1	0.86 (0.85)

Table S4: Capturing known-ligand binding sites at varying thresholds
Here, *n* designates the number of residues within a surface-critical site that overlap with known ligand-binding residues. For the calculations reported above and in the main text, this value is taken to be n=6 (because each surface-critical site typically has 10 residues, and never has more than 10 residues, this criterion enforces that a majority of surface-critical residues within a given site overlap with known ligand-binding residues in order to be counted as a site match). However, as this threshold is relaxed to lower n, the fraction of captured known ligand-binding sites improves rapidly, suggesting that surface-critical sites generally lie close to known ligand binding sites in many cases.

Protein (PDB, # residues)	Community Detection Method: GN | InfoMap			
	Modularity	# Comm.	# Critical Residues	% of GN critical residues which match those in Infomap (expected)
tRNA synthetase (1N78, 542)	0.71 | 0.68	14 | 25	47 | 109	0.28 (0.20)
Adenylate kinase (4AKE, 428)	0.73 | 0.70	11 | 20	39 | 82	0.90 (0.19)
Arginine Kinase (3JU5, 728)	0.72 | 0.69	12 | 28	41 | 142	0.22 (0.19)
Tyrosine Phosphatase (2HNP, 278)	0.59 | 0.59	7 | 15	27 | 70	0.26 (0.25)
Phosphoribosyltransferase (1XTT, 846)	0.72 | 0.68	9 | 32	36 | 174	0.22 (0.21)
cAMP-dep. PK (1J3H, 332)	0.66 | 0.64	11 | 19	36 | 78	0.33 (0.23)
Anthranilate synthase (1I7Q, 1418)	0.75 | 0.69	12 | 46	51 | 288	0.31 (0.20)
Malic enzyme (1EFK, 2212)	0.81 | 0.72	17 | 70	74 | 425	0.18 (0.19)
Threonine synthase (1E5X, 884)	0.73 | 0.69	13 | 36	43 | 192	0.28 (0.22)
G-6-P Deaminase (1CD5, 1596)	0.79 | 0.72	18 | 54	58 | 266	0.16 (0.17)
Phosphofructokinase (3PFK, 1276)	0.76 | 0.68	10 | 51	45 | 307	0.24 (0.24)
Tryptophan synthase (1BKS, 1294)	0.77 | 0.69	10 | 46	41 | 284	0.24 (0.22)
Means	***0.73 | 0.68***	***12.0 | 36.8***	***44.8 | 201.4***	***0.3***

Table S5: Comparing the two network module identification algorithms GN & Infomap
Though both GN (values to the left of "|" symbols throughout the table) and Infomap (values to the right) decompose networks to give similar modularity, the number of communities, and hence the number of critical residues connecting communities, is substantially larger when decomposing networks using Infomap than using GN.