Title:

Identifying allosteric hotspots with dynamics: application to inter- and intra-species conservation

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] Correspondence should be addressed to M.G. (pi@gersteinlab.org)

ABSTRACT

The rapidly growing volume of data being produced by next-generation sequencing initiatives is enabling more in-depth analyses of conservation than previously possible. Deep sequencing is uncovering disease loci and regions under selective constraint, despite the fact that intuitive biophysical reasons for such constraint are sometimes absent. Allostery may often provide the missing explanatory link. We use models of protein conformational change to identify allosteric residues by finding essential surface cavities and 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. Though fundamentally 3D-structural in nature, our analysis is computationally fast, thereby allowing us to run it across the PDB and to evaluate general properties of predicted allosteric residues. We find that these tend to be conserved over diverse evolutionary time scales. Finally, we highlight examples of allosteric residues that help explain poorly understood disease-associated variants.

INTRODUCTION

The ability to sequence large numbers of human genomes is providing a much deeper view into protein evolution than previously possible. 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 an integrated view of the context in which a protein and its residues function. This 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 be provided by studying the protein's dynamic behavior through the lens of the distinct functional and conformational states within an ensemble.

The underlying energetic landscape responsible for the relative distributions of alternative conformations is dynamic in nature: allosteric signals or other external changes may reconfigure and reshape the landscape, thereby shifting the relative populations of states within an ensemble (Tsai *et al*., 1999). Landscape theory thus provides the conceptual underpinnings necessary to describe how proteins change behavior and shape under changing conditions. A primary driving force behind the evolution of these landscapes is the need to efficiently regulate activity in response to changing cellular contexts, thereby making allostery and conformational change essential components of protein evolution.

Given the importance of allosteric regulation, as well as its role in imparting efficient functionality, several methods have been devised for the identification of likely allosteric residues. Conservation itself has been used, either in the context of conserved residues (Panjkovich and Daura, 2012), networks of co-evolving residues (Halabi *et al.*, 2009; Lee *et al.*, 2008; Lockless *et al.*, 1999; Reynolds *et al.*, 2011; Shulman *et al.*, 2004; Süel *et al.*, 2003), or local conservation in structure (Panjkovich and Daura, 2010). In related studies, both conservation and geometric-based searches for allosteric sites have been successfully applied to several systems (Capra *et al.*, 2009).

The concept of 'protein quakes' has been introduced to explain local conformational changes that are essential for global conformation transitions of functional importance (Ansari *et al.*, 1985; Miyashita *et al.*, 2003). These local changes cause strain within the protein that is relieved by subsequent relaxations (which are also termed functionally important motions) that terminate when the protein reaches the second equilibrium state. Such local perturbations often end with large conformational changes at the focal points of allosteric regulation, and these motions may be identified in a number of ways, including modified normal modes analysis (Miyashita *et al.*, 2003) or time-resolved X-ray scattering (Arnlund *et al.*, 2014).

In addition to conservation and geometry, protein dynamics have also been used to predict allosteric residues. Normal modes analysis has been used to examine the extent to which bound ligands interfere with low-frequency motions, thereby identifying potentially important residues at the surface (Ming and Wall, 2005; Mitternacht and Berezovsky, 2011; Panjkovich and Daura, 2012). Normal modes have also been used by the Bahar group to identify important subunits that act in a coherent manner for specific proteins (Chennubhotla and Bahar, 2006; Yang and Bahar, 2005). Rodgers *et al.* have applied normal modes to identify key residues in CRP/FNR transcription factors (Rodgers *et al*., 2013).

With the objective of identifying allosteric residues within the interior, molecular dynamics (MD) simulations and network analyses have been used to identify residues that may function as internal allosteric bottlenecks (Csermely *et al*., 2013; Gasper *et al*., 2012; Rousseau and Schymkowitz, 2005; Sethi *et al*., 2009; Vanwart *et al*., 2012). Ghosh *et al*. (2008) have taken a novel approach of combining MD and network principles to characterize allosterically important communication between domains in methionyl tRNA synthetase. In conjunction with NMR, Rivalta *et al.* have use MD and network analysis to identify important regions in imidazole glycerol phosphate synthase (Rivalta *et al*., 2012).

Though having provided valuable insights, many of these approaches have been 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). Here, we use models of protein conformational change to identify both surface and interior residues that may act as essential allosteric hotspots in a computationally tractable manner, thereby enabling high-throughput analysis. This framework directly incorporates information regarding 3D protein structure and dynamics, and it can be applied on a PDB-wide scale to proteins that exhibit conformational change. Throughout the PDB (Berman *et al*., 2000), the residues identified tend to be conserved both across species and amongst humans, and they 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, we make the software associated with this framework (termed STRESS, for STRucturally-identified ESSential residues) publically 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 often constitute the sources or sinks of allosteric signals, interior residues act to transmit such signals. We use models of protein conformational change to identify both classes of residues (Figure 1). Throughout, we term these potential allosteric residues at the surface and interior “surface-critical” and “interior-critical” residues, respectively.

In order to gauge the effectiveness of our approach, we identified and analyzed critical residues within a set of 12 well-studied canonical systems (see Figure S1, as well as Table S1 for rationale). We then apply this protocol on a large scale across hundreds of proteins for which crystal structures of alternative conformations are available.

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 (Figure 1A, top-left). The degree to which cavity occlusion by the ligand disrupts large-scale conformational change is used to assign a score to each cavity – sites at which ligand occlusion strongly interferes with conformational change earn high scores (Figure 1A, top-right), whereas shallow pockets (Figure 1A, bottom-left) or sites at which large-scale motions are largely unaffected (Figure 1A, bottom-right) earn lower scores. Further details are provided in SI Methods section 3.1-a.

This approach is a modified version of the binding leverage framework introduced by Mitternacht and Berezovsky (Mitternacht and Berezovsky, 2011). The main modifications implemented here 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. These modifications were implemented to provide a more selective set of sites; without them, a very large fraction of the protein surface would be occupied by critical sites (Figure S2C). Within our dataset of proteins exhibiting alternative conformations, we find that this modified approach results in an average of ~2 distinct sites per domain (Figure S2A; see Figure S2B for the distribution for distinct sites within entire complexes).

Within the canonical set of 12 proteins, we positively identify an average of 56% of the sites known to be directly involved in ligand or substrate binding (see Table 1, Figure S1, and SI Methods section 3.1-a-iv). 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 S2). 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 hitherto unfound 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 information flow ‘bottlenecks’ within the communication pathways between distant regions.

To identify such bottleneck residues, 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 Figure 1B and SI Methods section 3.1-b for details). Briefly, the network edges are first weighted by the degree of strength in 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 (Girvan *et al*., 2002). This formalism entails calculating the betweenness of each edge, where the betweenness of a given edge is defined as the number of shortest paths between all pairs of residues that pass through that edge (each path length is the sum of that path’s effective node-node distances assigned in the weighting scheme above). Each community identified 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 Figure 2.

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)

We have made the implementations for finding surface- and interior-critical residues available through a new software tool, STRESS, which may be accessed at stress.molmovdb.org (Figure 3A). Users may submit a PDB file or a PDB ID corresponding to a structure to be analyzed, and the output provided constitutes the set of identified critical residues.

Running times are minimized by using a scalable server architecture that runs on the Amazon cloud (Figure 3D). 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. In addition, the algorithmic implementation of our software is highly efficient, thereby obviating the need for long wait times. Relative to a naïve global Monte Carlo search implementation, local searches supported with hashing and additional algorithmic optimizations for computational efficiency reduce running times considerably (Figures 3B and 3C). A typical protein of ~500 residues takes only about 30 minutes on a 2.6GHz CPU.

High-Throughput Identification of Alternative Conformations

We use a generalized approach to systematically identify instances of alternative conformations throughout the PDB. We first perform multiple structure alignments (MSAs) across sequence-identical structures that are pre-filtered to ensure structural quality. We then use the resultant pairwise RMSD values to infer distinct conformational states (Figure S3; see also SI Methods section 3.2).

The distributions of the resultant numbers of conformations for domains and chains are given in Figures S3D and S3E, respectively, and an overview is given in Figure S3F. We note that the alternative conformations identified arise in an extremely diverse set of biological contexts, including conformational transitions that accompany ligand binding, protein-protein or protein-nucleic acid interactions, post-translational modifications, changes in oxidation or oligomerization states, etc. The dataset of alternative conformations identified is provided as a resource in File S1 (see also Figure S3G).

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

The large dataset of dynamic proteins culled throughout the PDB, coupled with the high algorithmic efficiency of our critical residue search implementation, provide a means of identifying and evaluating general properties of a large pool of critical residues. In particular, we use a variety of conservation metrics and data sources to measure the inter- and intra-species conservation of the residues within this pool. As discussed below, we find that both surface- (Figures 4A-D) and interior-critical residues (Figures 4E-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 predicted 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 (Figures 4B and 4F, respectively; note that negative conservation scores designate stronger conservation – see SI Methods section 3.3-a).

Leveraging Next-Generation Sequencing to Measure Conservation Amongst Humans

In addition to measuring inter-species conservation, we have also used fully sequenced human genomes and exomes to investigate conservation among human populations, as many constraints may be species-specific and active in more recent evolutionary history. Commonly used metrics for quantifying intra-species 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 section 3.3-b).

Non-synonymous single-nucleotide variants (SNVs) from the 1000 Genomes dataset (McVean *et al*., 2012) that intersect surface-critical residues tend to occur at lower DAF values than do SNVs that intersect non-critical residues (Figure 4C). Though this difference is not observed to be significant, the significance improves when examining the shift in DAF distributions, as evaluated with a KS test (p= 0.159, Figure S4A), and we point out only a limited number of proteins (thirty-two) for which these 1000 Genomes SNVs intersect with surface-critical sites. 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. In contrast to surface-critical residues, however, interior-critical residues intersect 1000 Genomes SNVs with significantly lower DAF values than do non-critical residues (Figure 4G; see also Figure S4B).

[[DC2MG(dec18): The paragraph added below was introduced after you suggested that we discuss the stats issue in the most recent annotated PDF. Another option might be to put this text in the Discussion instead of the Results, but I don't feel too strongly either way]] When analyzing human polymorphism data, a variety of statistical measures relating SNVs to selective constraint may be calculated, and the results obtained (along with their associated significance levels) are highly dependent on sample size. 1000 Genomes datasets are attractive partially because of their status as a well-established "blue chip" set of variants in human populations. However, given the relatively limited number of proteins that intersect with 1000 Genomes SNVs, we also analyzed the larger dataset provided by the Exome Aggregation Consortium (ExAC) (Exome Aggregation Consortium, 2015). Though this dataset has been released much more recently (and is consequently not yet as well established as 1000 Genomes), ExAC provides sequence data from more than 60,000 individuals, and samples are sequenced at much higher coverage, thereby ensuring better data quality. This larger dataset enables us to more easily examine trends in the data as they relate to critical and non-critical residues.

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 Figure 4D. 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 (Figure S5A, KS test p=9.49e-2).

Interior-critical residues exhibit significantly lower MAF values than do non-critical residues in the same set of proteins. MAF distributions for interior- and non-critical residues are given in Figure 4H (see also Figure S5B).

In addition to analyzing overall allele frequency distributions, we 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 rare (i.e., low-DAF or low-MAF) non-synonymous SNVs to the number of all non-synonymous SNVs in a given protein annotation (such as all surface-critical residues of the protein, for example; see SI Methods section 3.3-b). A higher fraction is interpreted as a proxy for greater conservation (Khurana *et al*., 2013; Sethi *et al*., 2015). 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 on critical residues (See Figure 5).

Comparisons Between Different Models of Protein Motions

The identification of surface- and interior-critical residues entails using sets of vectors (on each protein residue) to describe conformational change. Notably, our framework enables one to determine these vectors in multiple ways. Conformational changes may be modeled using vectors connecting residues in crystal structures from alternative conformations. We term this approach “ACT”, for “absolute conformational transitions” (see SI Methods section 3.2-c). The crystal structures of such paired conformations may be obtained using the framework discussed above. The protein motions may also be inferred from anisotropic network models (ANMs) (Atilgan *et al*., 2001). ANMs entail modeling interacting residues as nodes linked by flexible springs, in a manner similar to elastic network models (Fuglebakk *et al*., 2015; Tirion, 1996) or normal modes analysis (Figure 1B). ANMs are not only simple and straightforward to apply on a database scale, but unlike using alternative crystal structures, the motion vectors inferred may be generated using a single structure.

We find that modeling conformational change using vectors from either ACTs or ANMs gives the same general trends in terms of the disparities in conservation between critical and non-critical residues. Our framework is thus general with respect to how the motion vectors are obtained (see Figure 6 and SI Methods section 3.2-c for further details).

Critical Residues in the Context of Human Disease Variants

Directly related to conservation is confidence with which an SNV is believed to be disease-associated. SIFT (Ng and Henikoff, 2001) and PolyPhen (Adzhubei *et al*., 2010) are two tools for predicting SNV deleteriousness. ExAC SNVs that intersect critical residues exhibit significantly higher PolyPhen scores relative to non-critical residues, suggesting the potentially higher disease susceptibility at critical residues (Figure S6). Significant disparities were not observed in SIFT scores (Figure S7).

Using HGMD (Stenson *et al*., 2014) and ClinVar (Landrum *et al*., 2014), we identify proteins with critical residues that coincide with disease-associated SNVs (Figure 7A 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 (Figure 7). SNVs in FGFR have been linked to craniofacial defects. Dotted lines in Figure 7B highlight poorly understood disease SNVs that coincide with critical residues. In addition, we identify Y328 as a surface-critical residue, which coincides with a disease-associated SNV from HGMD, despite the lack of confident predictions of deleteriousness by several widely used tools for predicting disease-associated SNVs, including PolyPhen (Adzhubei *et al*., 2010), SIFT (Ng and Henikoff, 2001), and SNPs&GO (Calabrese *et al*., 2009). 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 fully 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 to 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 been unknown, but for which allosteric mechanisms provide a reasonable 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 through 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 (Kornev and Taylor, 2015). 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.

Despite these limitations in MD and NMR, allosteric mechanisms and signaling pathways may be conserved across many different but related proteins within the same family, suggesting that such computationally- or labor-intensive approaches for all proteins may not be entirely essential. Flock *et al*. have carefully demonstrated that the allosteric mechanisms responsible for regulating G proteins through GPCRs tend to be conserved (Flock *et al*., 2015). Investigations into representative families have also been enlightening in other contexts. In one of the early studies employing network analysis, del Sol *et al*. conduct a detailed study of several allosteric protein families (including GPCRs) to demonstrate that residues important for maintaining the integrity of short paths within residue contact networks are essential to enabling signal transmission between distant sites (del Sol *et al.*, 2006). Another notable result in the same work is that these key residues (which match experimental results) may become redistributed when the protein undergoes conformational change, thereby changing optimal communication routes as a means of conferring different regulatory properties.

There are several notable implications of our dynamics-based analysis across a database of proteins. 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.

It is only by analyzing a large dataset of proteins can one investigate general trends in predicted allosteric residues. In addition, the implementation detailed here enables one to match structural features with the high-throughput data generated through deep sequencing initiatives, which are providing an unprecedented window into conservation patterns, many of which may be human-specific.

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 species evolve in various evolutionary contexts and at different rates, 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 particular, selective constraints within human populations are particularly relevant to understanding human disease. Formalisms for analyzing large structural and sequence datasets will become increasingly important in the context of human health. We anticipate that the framework and formalisms detailed here, along with the accompanying web tool we have introduced, will help to further motivate future studies along these directions.

METHODS

An overview of the framework for finding surface- and interior-critical residues is given in Figure 1. Figure S3 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 protocols 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

Adzhubei, I. Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods. *7*, 248–249

Ansari, A., Berendzen, J., Bowne, S., Frauenfelder, H., Iben, I.E.T., Sauke, T.B., Shyamsunder, E., and Young, R.D. (1985). Protein states and protein quakes. Proc. Natl. Acad. Sci. U.S.A. 82, 5000–5004.

Arnlund, D., Johansson, L.C., Wickstrand, C., Barty, A., Williams, G.J., Malmerberg, E., Davidsson, J., Milathianaki, D., DePonte, D.P., Shoeman, R.L., *et al*. (2014). Visualizing a protein quake with time-resolved X-ray scattering at a free-electron laser. Nat. Methods. *11*, 923–6.

Atilgan, A.R., Durell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O., and Bahar, I. (2001). Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. Biophys. J. *80*, 505–515.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids

Res. *28*, 235–242.

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P.L. and Casadio, R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum. Mutat. *30*, 1237–1244.

Exome Aggregation Consortium, Lek, M., Karczewski, K., Minikel, E., Samocha, K., Banks, E., Fennell, T., O'Donnell-Luria, A., Ware, J., Hill, A., et al. (2015). Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv. 030338

Capra, J.A., Laskowski, R.A., Thornton, J.M., Singh, M. and Funkhouser, T.A. (2009). Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput. Biol. *5*, e1000585.

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

del Sol, A., Fujihashi, H., Amoros, D., and Nussinov, R. (2006). Residues crucial for maintaining short paths in network communication mediate signaling in proteins. Mol. Syst. Biol. 2(1).

Csermely, P., Korcsmáros, T., Kiss, H.J.M., London, G., and Nussinov, R. (2013). Structure and dynamics of molecular networks: A novel paradigm of drug discovery. Pharmacol. Ther. *138*, 333–408.

Flock, T., Ravarani, C.N.J., Sun, D., Venkatakrishnan, A.J., Kayikci, M., Tate, C.G., Veprintsev, D.B. and Babu, M.M. (2015). Universal allosteric mechanism for Gα activation by GPCRs. Nature *524*, 173–179.

Fuglebakk, E., Tiwari, S.P., and Reuter, N. (2015). Comparing the intrinsic dynamics of multiple protein structures using elastic network models. Biochim. Biophys. Acta -

Gen. Subj. *1850*, 911–922.

Gasper, P.M., Fuglestad, B., Komives, E.A., Markwick, P.R.L., and McCammon, J.A. (2012). Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities. Proc. Natl. Acad. Sci. U. S. A. *109*, 21216–22.

Ghosh, A., and Vishveshwara, S. (2008). Variations in Clique and Community Patterns in Protein Structures during Allosteric Communication: Investigation of Dynamically Equilibrated Structures of Methionyl tRNA Synthetase Complexes. Biochemistry. *47*, 11398-11407.

Girvan, M., Girvan, M., Newman, M.E.J., and Newman, M.E.J. (2002). Community structure in social and biological networks. Proc. Natl. Acad. Sci. U. S. A. *99*, 7821–7826.

Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009). Protein Sectors: Evolutionary Units of Three-Dimensional Structure. Cell *138*, 774–786.

Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., *et al*. (2013). Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. Science. *342*, 1235587–1235587.

Kornev, A.P. and Taylor, S.S. (2015). Dynamics-Driven Allostery in Protein Kinases. Trends Biochem. Sci. *xx*, 1–20.

Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. *42*, D980–5.

Lee, J., Natarajan, M., Nashine, V.C., Socolich, M., Vo, T., Russ, W.P., Benkovic, S.J., and Ranganathan, R. (2008). Surface Sites for Engineering Allosteric Control in Proteins. Science *322*, 438-442.

Lockless, S.W., Ranganathan, R., Kukic, P., Mirabello, C., Tradigo, G., Walsh, I., Veltri, P., Pollastri, G., Socolich, M., Lockless, S.W., *et al*. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. BMC Bioinformatics *15*, 295–299.

McVean, G.A., Altshuler (Co-Chair), D.M., Durbin (Co-Chair), R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., *et al*. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Ming, D. and Wall, M.E. (2005). Quantifying allosteric effects in proteins. Proteins *59*, 697–707.

Mitternacht, S. and Berezovsky, I.N. (2011). Binding leverage as a molecular basis for allosteric regulation. PLoS Comput. Biol. *7*, e1002148.

Miyashita, O., Onuchic, J.N., and Wolynes, P.G. (2003). Nonlinear elasticity, protein quakes, and the energy landscapes of functional transitions in proteins. Proc. Natl. Acad. Sci. *100*, 12570–12575.

Ng, P.C. and Henikoff, S. (2001). Predicting Deleterious Amino Acid Substitutions. Genome Res. *11*, 863–874.

Panjkovich, A. and Daura, X. (2012). Exploiting protein flexibility to predict the location of allosteric sites. BMC Bioinformatics *13*, 273.

Panjkovich, A. and 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.

Reynolds, K.A., McLaughlin, R.N., and Ranganathan, R. (2011). Hot Spots for Allosteric Regulation on Protein Surfaces. Cell *147*, 1564–1575.

Rivalta, I., Sultan, M.M., Lee, N.-S., Manley, G. a., Loria, J.P., and Batista, V.S. (2012). PNAS Plus: Allosteric pathways in imidazole glycerol phosphate synthase. Proc. Natl. Acad. Sci. *109*, E1428–E1436.

Rodgers, T.L., Townsend, P.D., Burnell, D., Jones, M.L., Richards, S.A., McLeish, T.C.B., Pohl, E., Wilson, M.R., and Cann, M.J. (2013). Modulation of Global Low-Frequency Motions Underlies Allosteric Regulation: Demonstration in CRP/FNR Family Transcription Factors. PLoS Biol. *11*, e1001651.

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

Sethi, A., Eargle, J., Black, A.A., and Luthey-Schulten, Z. (2009). Dynamical networks in tRNA:protein complexes. Proc. Natl. Acad. Sci. U. S. A. *106*, 6620–5.

Sethi, A., Clarke, D., Chen, J., Kumar, S., Galeev, T.R., Regan, L., and Gerstein, M. (2015). Reads meet rotamers: structural biology in the age of deep sequencing. Curr. Opin. Struct. Biol. *35*, 125-134.

Shulman, A.I., Larson, C., Mangelsdorf, D.J., and Ranganathan, R. (2004). Structural determinants of allosteric ligand activation in RXR heterodimers. Cell *116*, 417–429.

Stenson, P.D., Mort, M., Ball, E. V., Shaw, K., Phillips, A.D., and Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum. Genet. *133*, 1–9.

Süel, G.M., Lockless, S.W., Wall, M.A., and Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat. Struct. Biol. *10*, 59–69.

Tirion, M.M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. Phys. Rev. Lett. *77*, 1905–1908.

Tsai, C., Ma, B. and Nussinov, R. (1999). Folding and binding cascades: Shifts in energy landscapes. Proc. Natl. Acad. Sci. U. S. A. *96*, 9970–9972.

Vanwart, A.T., Eargle, J., Luthey-Schulten, Z., and Amaro, R.E. (2012). Exploring residue component contributions to dynamical network models of allostery. J. Chem. Theory Comput. *8*, 2949–2961.

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

CAPTIONS

Figure 1. 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 effect 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.

Figure 2. 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 in SI Methods section 3.1-b. 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.

Figure 3. STRESS web server front page, running times, and server architecture. (A) 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. (B) Running times are shown for systems of various sizes. Shown in red are the running times without optimizing for speed, and green shows running times with algorithmic optimization. (C) 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. (D) 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).

Figure 4. 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 (B), mean inter-species conservation scores for surface-critical sets are -0.131, whereas non-critical residue sets with the same degree of burial have a mean score of +0.059 (p < 2.2e-16). In (F), mean inter-species conservation scores for interior-critical sets are -0.179, whereas non-critical residue sets with the same degree of burial have a mean score of -0.102 (p=3.67e-11). 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). 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.

Figure 5. Critical residues are shown to be more conserved, as measured by the fraction of rare alleles. Protein regions with high fractions of *rare* variants are believed to be more sensitive to sequence variants than other regions, thereby explaining why such variants occur infrequently in the population. Panels *(A)* and *(C)* show distributions for rare (low DAF) non-synonymous SNVs (taken from the 1000 Genomes dataset) in which the critical residues are defined to be the surface-critical *(A)* and interior-critical *(C)* residues. Panels *(B)* and *(D)* show distributions for rare (low MAF) non-synonymous SNVs (taken from the ExAC dataset) in which the critical residues are defined to be the surface-critical *(B)* and interior-critical *(D)* residues. 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. We consider all structures such that at least one critical and at least one non-critical residue intersect a non-synonymous SNV. Panels *(A), (B), (C),* and *(D)* represent data from 31, 90, 32, and 84 structures, respectively.

Figure 6. Modeling protein conformational change through a direct use of crystal structures from alternative conformations using absolute conformational transitions (ACT). *(A)* Distributions (155 structures) of the mean conservation scores on surface-critical (red) and non-critical residues with the same degree of burial (blue). *(B)* 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.

Figure 7. Potential allosteric residues add a layer of annotation to structures in the context of disease-associated SNVs. The structure shown (*A*) is that of the fibroblast growth-factor receptor (FGFR) in VMD Surf rendering, with HGMD SNVs shown in orange, bound to FGF2, in ribbon rendering (PDB 1IIL). (*B*) A linear representation of structural annotation for FGFR. Dotted lines highlight loci which correspond to HGMD sites that coincide with critical residues, but for which other annotations fail to coincide. Deeply-buried residues are defined to be those that exhibit a relative solvent-exposed surface area of 5% or less, and binding site residues are defined as those for which at least one heavy atom falls within 4.5 Angstroms of any heavy atom in the binding partner (heparin-binding growth factor 2). The loci of PTM sites were taken from UniProt (accession P21802).

Table 1. Statistics on the surfaces of *apo* structures within the canonical set of proteins

For each *apo* structure within the canonical set of proteins, statistics relating surface-critical sites to known ligand-binding sites are reported. The surface of a given structure is defined to be the set of all residues that have a relative solvent accessibility of at least 50%, where relative solvent accessibility is evaluated using all heavy atoms in both the main-chain and side-chain of a given residue. Mean values are given in the bottom row.

NACCESS is used to calculate relative solvent accessibility (Hubbard and Thornton, 1993) . *Column 1*: PDB IDs for each structure; *Column 2*: among these surface residues, the fraction that constitute surface-critical residues; *Column 3*: among surface residues, the fraction that constitute known ligand-binding residues (known ligand-binding residues are taken to be those within 4.5 Angstroms of the ligand in the *holo* structure; Table S1); *Column 4*: the Jaccard similarity between the sets of residues represented in columns 2 and 3 (i.e., surface-critical and known-ligand binding residues), where values given in parentheses represent the expected Jaccard similarity, given a null model in which surface-critical and ligand-binding residues are randomly distributed throughout the surface (for each structure, 10,000 simulations are performed to produce random distributions, and the expected values reported here constitute the mean Jaccard similarity among the 10,000 simulations for each structure); *Column 5*: the number of distinct surface-critical sites identified in each structure; *Column 6*: the number of known ligand-binding sites in each structure; *Column 7*: the number of known ligand-binding sites which are positively identified within the set of surface-critical sites, where a positive match occurs if a majority of the residues in a surface-critical site coincide with the known ligand-binding site; *Column 8*: The fraction of ligand-binding sites captured is simply the ratio of the values in column 7 to those in column 6.