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)

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: Predicting Allosteric Hotspots Using Dynamics-Based Formalisms with Sequence Analyses Across Diverse Evolutionary Timescales ... [1]

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: [a]Department

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: [b]Program

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: [c]Department

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: [d]Department

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: [e]Yale

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: [f]Integrated

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: (pi@gersteinlab.org)

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Line spacing: 1.5 lines

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 missing. 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.

ABSENT

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: unavailable

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: predicting

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: or

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: this software

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: the

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: , which

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: long and short

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: We

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: in which

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: can

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 between 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

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: .

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: amongst

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: a more

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: integrated

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: often

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: considering

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: and

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

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

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: the

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: of allostery

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: A number of methods employing support vector machines have also been described (Huang and Schroeder, 2006; Huang et al., 2013). Normal modes analysis, coupled with ligands of varying size, have

DECLAN CLARKE 12/13/15 2:56 PM
Moved down [1]: 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).

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: The concept of ‘protein quakes’ has been introduced to explain local regions of proteins that are essential for conformation transitions (Miyashita et al., 2003). A protein may relieve the strain of a high-energy configuration by local structural changes. Such local changes often occur at the focal points of allosteric regulation, and these regions may be identified in a number of ways, including modified normal modes analysis (Miyashita et al., 2003) or time-resolved X-ray scattering (Arnlund et al., 2014).

DECLAN CLARKE 12/13/15 2:56 PM
Moved (insertion) [1]

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

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: Molecular

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: interior

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: Along similar lines,

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: inter-domain

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: (Ghosh et al., 2008).

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: may be

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: Using

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: , we

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: regions

dynamics, and it can be applied on a PDB-wide scale to proteins that exhibit conformational change (Berman *et al.*, 2000). The residues identified tend to be conserved both across species and between 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 may 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.

Critical residues are identified and analyzed within a set of 12 well-studied canonical systems (see Figure S1, as well as Table S1 for rationale), and they are then

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: is

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: throughout the PDB (Berman et al., 2000)

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: .

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: relatively greater conservation of the

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: (

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: amongst

DECLAN CLARKE 12/13/15 2:56 PM
Deleted:)

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: our

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: is made

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Line spacing: 1.5 lines, Tabs: 2.3", Left

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: in an attempt

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: first

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: in

TO GAUGE THE EFFECTIVENESS OF OUR APPROACH :)

investigated on a large scale across hundreds of proteins for which crystal structures of alternative conformations are available.

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: identified

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: distinct

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.

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: interfere

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, an exceedingly large fraction of the protein surface would be captured (Figure 2C). 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 2A; see Figure 2B for the distribution for distinct sites within entire complexes).

WD

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: . Without

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: We

DECLAN CLARKE 12/13/15 2:56 PM
Deleted:). The

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: is given in Figure 2B.

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 [[DC2MG(12/11): I actually don't know if I fully agree with this change that was introduced: when we talk about latent allosteric sites, the thing this was previously unfound is not the *motions* themselves, but rather the *pockets* which were not previously known to disrupt already-known motions. We can discuss during P2 struct]].

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

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Indent: First line: 0.5", Tabs: 3.56", Left

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: large-scale motions.

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Arial, 11 pt, Bold

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: distal

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: 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.

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: bottlenecks

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). 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 Figures S2.

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)

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

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.

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

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

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

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: The

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: both

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: have been made

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: to the scientific community

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: specify a PDB

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: A typical protein of ~500 residues takes only about 30 minutes on a 2.6GHz CPU.

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: also

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: Pronounced conformational change is an essential assumption within our framework for identifying potential allosteric residues.

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: within

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: proteins

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 evaluating general properties within the large pool of critical residues identified. 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

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Indent: First line: 0"

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: number

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: of these

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: on a large scale.

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: measure their conservation, as evaluated both over long (inter-species) and short (intra-human) evolutionary timescales. Using

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: data

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: critical

many of the otherwise poorly understood regions of proteins, but they also reinforce the functional importance of the residues predicted to be allosteric.

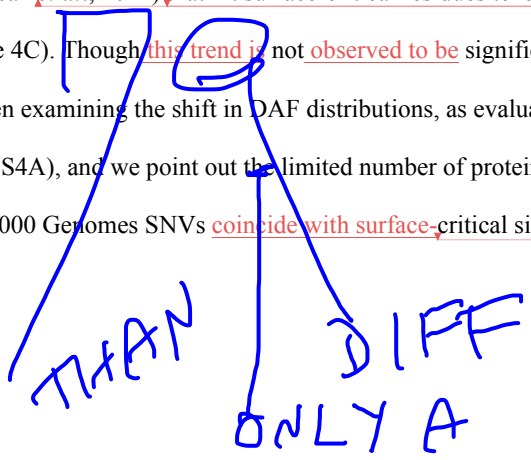
Conservation Across Species

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 Between 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 hit surface-critical residues tend to occur at lower DAF values (Figure 4C). Though this trend 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 the limited number of proteins (thirty-two) for which these 1000 Genomes SNVs coincide with surface-critical sites. Furthermore, the



DECLAN CLARKE 12/13/15 2:56 PM
Deleted: believed

DECLAN CLARKE 12/13/15 2:56 PM
Deleted:). Surface-critical residue sets have a mean conservation score (i.e., ConSurf score, see SI Methods section 3.3-a) of -0.131, whereas non-critical residue sets with the same degree of burial have a mean score of +0.059 ($p < 2.2e-16$;

DECLAN CLARKE 12/13/15 2:56 PM
Deleted:). 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 ($p=3.67e-11$

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: Measures of Conservation Amongst Humans from

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: We may

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: use

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: human

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: In this context, commonly

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: evaluating

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: We find that

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: single-nucleotide variants (SNVs)

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: in

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: hit these

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 are hit by 1000 Genomes SNVs with significantly lower DAF values than non-critical residues (Figure 4G; see also Figure S4B).

Given the limited number of proteins to be hit by 1000 Genomes SNVs, we also analyzed the larger dataset provided by the Exome Aggregation Consortium (ExAC, Cambridge MA 2015). ExAC provides sequence data from more than 60,000 individuals, and samples are sequenced at much higher coverage, thereby ensuring better data quality. 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

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: However 1000 Genomes SNVs tend to hit

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: relatively small

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: data

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: for many

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: the ExAC sequencing itself is performed

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: . Thus, using

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: one may

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: (i.e., rare

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). 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.

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

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: in

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: pairs of corresponding

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: (we

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: ";

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: , and we thus use ANMs as our primary means of inferring motions

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: Using

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: give

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: results

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 hitting 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.

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: This method

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: defined

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: the concept of SNV deleteriousness: changes in amino acid composition at specific loci may be more or less likely

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: result in

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: such effects, and we evaluated these predictions for critical and non-critical residues hit by SNVs in

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: .

DECLAN CLARKE 12/13/15 2:56 PM
Deleted:), though such significant

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: HGDM

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

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 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 through dynamic behavior. It is only by considering protein motions and changes in these motions can a fuller understanding of

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Line spacing: 1.5 lines

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: in the context of

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: previously

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: unavailable

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: in the context of

564 allosteric regulation be realized. As such, MD and NMR are some of the most common
565 means of studying allostery and dynamic behavior (Kornev and Taylor, 2015). However,
566 these methods have limitations when studying large and diverse protein datasets. MD is
567 computationally expensive and impractical when studying large numbers of proteins.
568 NMR structure determination is extremely labor-intensive and better suited to certain
569 classes of structures or dynamics. In addition, NMR structures constitute a relatively
570 small fraction of structures currently available.

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

585 There are several notable implications of our dynamics-based analysis across a
586 database of proteins. Relative to sequence data, allostery and dynamic behavior are far

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: If allosteric mechanisms are similarly shared within other protein families, a detailed analysis with methods such as MD or NMR on one member of a family may help to elucidate the allosteric behavior for other members. Nevertheless, the degree to which these mechanisms are indeed conserved within other groups of proteins is currently unclear, so homology-based predictions of allosteric mechanisms are still not readily available.

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Font:Italic

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: ,

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: Notably, many of the key sites identified correspond to residues that had been experimentally determined to be important for allostery.

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: in different conformations

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: -scale analysis.

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 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 particular, selective constraints within human populations are particularly relevant to understanding human disease. Formalisms for analyzing large structural and

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: 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.

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: We emphasize that it

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: applying this framework over a database of many

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: search for significant disparities in conservation between sites believed to be important in allostery and the rest of the protein. Such

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: may not be apparent when studying a small number or specific classes of proteins. To our knowledge, this is the first study in which the conservation of potential allosteric sites has been measured across a large database of proteins. ... [2]

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: . Full human genomes and exomes are being sequenced at an increasing pace, thereby

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: that can be human-specific or active over short evolutionary timescales. These patterns increasingly serve as detailed signatures of selective constraints which may not only be missing in cross-species comparisons, but are also sometimes difficult to rationalize using static representations of protein structures alone

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Indent: First line: 0.5"

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: addition, intra-species

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

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: disease. Finally, we

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: our newly developed software

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: will prove to be of great value in enabling investigators to study allostery in diverse contexts.

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Line spacing: 1.5 lines

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Indent: First line: 0.5"

DECLAN CLARKE 12/13/15 2:56 PM
Deleted:

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Line spacing: 1.5 lines

DECLAN CLARKE 12/13/15 2:56 PM
Deleted:

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

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 (ExAC). (2015) Cambridge, MA. http://exac.broadinstitute.org.

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

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Indent: Left: 0", Hanging: 0.33", Line spacing: 1.5 lines

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.

733 Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009). Protein Sectors:
734 Evolutionary Units of Three-Dimensional Structure. Cell *138*, 774–786.
735 Huang, B., and Schroeder, M. (2006). LIGSITEcsc: predicting ligand binding sites using
736 the Connolly surface and degree of conservation. BMC Struct. Biol. *6*, 19.
737 Huang, W., Lu, S., Huang, Z., Liu, X., Mou, L., Luo, Y., Zhao, Y., Liu, Y., Chen, Z.,
738 Hou, T., et al. (2013). Allosite: A method for predicting allosteric sites.
739 Bioinformatics *29*, 2357–2359.
740 Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A.,
741 Lochovsky, L., Chen, J., Harmanci, A., et al. (2013). Integrative Annotation of
742 Variants from 1092 Humans: Application to Cancer Genomics. Science. *342*,
743 1235587–1235587.
744 Kornev, A.P. and Taylor, S.S. (2015). Dynamics-Driven Allostery in Protein Kinases.
745 Trends Biochem. Sci. *xx*, 1–20.
746 Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and
747 Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence
748 variation and human phenotype. Nucleic Acids Res. *42*, D980–5.
749 Lee, J., Natarajan, M., Nashine, V.C., Socolich, M., Vo, T., Russ, W.P., Benkovic, S.J.,
750 and Ranganathan, R. (2008). Surface Sites for Engineering Allosteric Control in
751 Proteins. Science *322*, 438-442.
752 Lockless, S.W., Ranganathan, R., Kukic, P., Mirabello, C., Tradigo, G., Walsh, I., Veltri,
753 P., Pollastri, G., Socolich, M., Lockless, S.W., et al. (1999). Evolutionarily
754 conserved pathways of energetic connectivity in protein families. BMC
755 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.

779 Rodgers, T.L., Townsend, P.D., Burnell, D., Jones, M.L., Richards, S.A., McLeish,
780 T.C.B., Pohl, E., Wilson, M.R., and Cann, M.J. (2013). Modulation of Global Low-
781 Frequency Motions Underlies Allosteric Regulation: Demonstration in CRP/FNR
782 Family Transcription Factors. PLoS Biol. *11*, e1001651.
783 Rousseau, F. and Schymkowitz, J. (2005). A systems biology perspective on protein
784 structural dynamics and signal transduction. Curr. Opin. Struct. Biol. *15*, 23–30.
785 Sethi, A., Eargle, J., Black, A.A., and Luthey-Schulten, Z. (2009). Dynamical networks
786 in tRNA:protein complexes. Proc. Natl. Acad. Sci. U. S. A. *106*, 6620–5.
787 Shulman, A.I., Larson, C., Mangelsdorf, D.J., and Ranganathan, R. (2004). Structural
788 determinants of allosteric ligand activation in RXR heterodimers. Cell *116*, 417–
789 429.
790 Stenson, P.D., Mort, M., Ball, E. V., Shaw, K., Phillips, A.D., and Cooper, D.N. (2014).
791 The Human Gene Mutation Database: building a comprehensive mutation repository
792 for clinical and molecular genetics, diagnostic testing and personalized genomic
793 medicine. Hum. Genet. *133*, 1–9.
794 Süel, G.M., Lockless, S.W., Wall, M.A., and Ranganathan, R. (2003). Evolutionarily
795 conserved networks of residues mediate allosteric communication in proteins. Nat.
796 Struct. Biol. *10*, 59–69.
797 Tirion, M.M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-
798 Parameter, Atomic Analysis. Phys. Rev. Lett. *77*, 1905–1908.
799 Tsai, C., Ma, B. and Nussinov, R. (1999). Folding and binding cascades: Shifts in energy
800 landscapes. Proc. Natl. Acad. Sci. U. S. A. *96*, 9970–9972.
801 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

DECLAN CLARKE 12/13/15 2:56 PM
Formatted: Line spacing: 1.5 lines

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. Summary statistics for surface-critical sites. The distributions of the numbers of surface-critical sites per domain and per complex are given in (*A*) and (*B*), respectively. Panel (C) gives the distributions of the number of surface-critical sites per

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

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

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: :

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: . Performing local searching supported with hashing

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: implementing additional algorithmic optimizations for computational efficiency reduce

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: considerably (in green), relative to a more naïve approach without

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: (in red).

DECLAN CLARKE 12/13/15 2:56 PM
Deleted:

858 3PFK) are given in (*A*) and (*E*), respectively. Distributions of cross-species conservation
859 scores, 1000 Genomes SNV DAF averages, and ExAC SNV MAF averages for surface-
860 and non-critical residue sets are given in (*B*), (*C*), and (*D*), respectively. The same
861 distributions corresponding to interior- and non-critical residue sets are given in (F), (G),
862 and (H), respectively. In (B), mean inter-species conservation scores for surface-critical
863 sets are -0.131, whereas non-critical residue sets with the same degree of burial have a
864 mean score of +0.059 ($p < 2.2e-16$). In (F), mean inter-species conservation scores for
865 interior-critical sets are -0.179, whereas non-critical residue sets with the same degree of
866 burial have a mean score of -0.102 (p=3.67e-11). In (C), means for surface- and non-
867 critical sets are 9.10e-4 and 8.34e-4, respectively (p=0.309); corresponding means in (*D*)
868 are 4.09e-04 and 2.26e-04, respectively (p=1.49e-3). In (*G*), means for interior- and non-
869 critical sets are 2.82e-4 and 3.12e-3, respectively (p=1.80e-05); corresponding means in
870 (*H*) are 3.08e-05 and 3.27e-04, respectively (p=7.98e-09). N = 421, 32, 84, 517, 31, and
871 90 structures for panels B, C, D, F, G, and H, respectively. P-values are based on
872 Wilcoxon-rank sum tests. See SI Methods for further details.
873
874 **Figure 5. Critical residues are shown to be more conserved, as measured by the**
875 **fraction of rare alleles.** Protein regions with high fractions of *rare* variants are believed
876 to be more sensitive to sequence variants than other regions, thereby explaining why such
877 variants occur infrequently in the population. Panels *(A)* and *(C)* show distributions for
878 rare (low DAF) non-synonymous SNVs (taken from the 1000 Genomes dataset) in which
879 the critical residues are defined to be the surface-critical *(A)* and interior-critical *(C)*
880 residues. Panels *(B)* and *(D)* show distributions for rare (low MAF) non-synonymous

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: Statistics for panels (*B*) and (*F*) are given in the main text.

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: :

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 are hit by 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.

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: :

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.

908 Deeply-buried residues are defined to be those that exhibit a relative solvent-exposed
909 surface area of 5% or less, and binding site residues are defined as those for which at
910 least one heavy atom falls within 4.5 Angstroms of any heavy atom in the binding partner
911 (heparin-binding growth factor 2). The loci of PTM sites were taken from UniProt
912 (accession P21802).
913
914 **Table 1. Statistics on the surfaces of *apo* structures within the canonical set of**
915 **proteins**
916 For each *apo* structure within the canonical set of proteins, statistics relating surface-
917 critical sites to known ligand-binding sites are reported. The surface of a given structure
918 is defined to be the set of all residues that have a relative solvent accessibility of at least
919 50%, where relative solvent accessibility is evaluated using all heavy atoms in both the
920 main-chain and side-chain of a given residue. Mean values are given in the bottom row.
921 NACCESS is used to calculate relative solvent accessibility (Hubbard and Thornton,
922 1993) . *Column 1*: PDB IDs for each structure; *Column 2*: among these surface residues,
923 the fraction that constitute surface-critical residues; *Column 3*: among surface residues,
924 the fraction that constitute known ligand-binding residues (known ligand-binding
925 residues are taken to be those within 4.5 Angstroms of the ligand in the *holo* structure;
926 Table S1); *Column 4*: the Jaccard similarity between the sets of residues represented in
927 columns 2 and 3 (i.e., surface-critical and known-ligand binding residues), where values
928 given in parentheses represent the expected Jaccard similarity, given a null model in
929 which surface-critical and ligand-binding residues are randomly distributed throughout
930 the surface (for each structure, 10,000 simulations are performed to produce random

DECLAN CLARKE 12/13/15 2:56 PM
Deleted: :

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.

Page 1: [1] Deleted DECLAN CLARKE 12/13/15 2:56 PM

Predicting Allosteric Hotspots Using Dynamics-Based Formalisms with Sequence Analyses Across Diverse Evolutionary Timescales

Page 18: [2] Deleted DECLAN CLARKE 12/13/15 2:56 PM

may not be apparent when studying a small number or specific classes of proteins.

To our knowledge, this is the first study in which the conservation of potential allosteric sites has been measured across a large database of proteins.

The ability to leverage our framework in a high-throughput manner also better