# ABSTRACT

The rapidly growing volumes of data being produced by next-generation sequencing initiatives are enabling more in-depth analyses of protein conservation than previously possible. Deep sequencing is uncovering disease loci and protein regions under selective constraint, despite the fact that intuitive biophysical mechanisms responsible for such constraints are sometimes lacking (such as the need to engage in protein-protein interactions or undergo post-translational modifications). Allosteric hotspots may often provide the missing conceptual link, and we use models of protein conformational change to identify such residues. In particular, models of conformational change are used to predict allosteric cavities on the surface, as well as allosteric residues which can function as information flow bottlenecks within the interior. A web server has been developed to enable users to perform this analysis on their own proteins of interest, and we note that this approach is both computationally tractable and fundamentally structural in nature – conformational change and topology are directly included in the search for allosteric residues. Finally, by developing a method for automatically culling instances of alternative conformations throughout the PDB, allosteric hotspot predictions are made on a database-level scale, and downstream analyses of the predicted allosteric residues reveal that they tend to be conserved across both long and short evolutionary time scales.

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** The identification of potentially allosteric residues is essential to understanding protein function, and allostery provides a potential means to understanding many of the conservation patterns observed in sequence space. Meanwhile, the

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** . Potentially allosteric residues

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** develop

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** essential patches

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** identified

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** essential

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** Though most existing approaches entail computationally expensive methods (such as MD) or rely on less direct measures (such as sequence features), our framework is simultaneously

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** essential residues. Downstream analyses of these residues reveal that they tend to be conserved across both long and short evolutionary time scales (i.e., across species and among modern-day humans). We also introduce a web server to enable users to perform this analysis on their own proteins of interest.

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** our analysis is applied

# INTRODUCTION

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

in understanding these regions may often be provided by considering the protein's dynamic behavior and distinct functional states within an ensemble.

In addition to the multiple conformations exhibited by a given protein, the underlying energetic landscape itself 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 regulate activity and efficiency in response to changing cellular contexts, thereby making allostery and conformational change essential components of protein evolution.

An allosteric mechanism may involve the modulation of large-scale motions upon binding of an effector ligand, resulting in conformational changes at distant surface sites. Such motions may also affect patterns of communication between residues, and internal residues essential to the integrity of these communication networks constitute bottlenecks.

Given the importance of allosteric regulation, as well as the role of allostery in imparting efficient functionality, several methods have been devised for the prediction of allosteric residues. Conservation has been used, either in the context of conserved residues (Panjkovich and Daura, 2012), networks of co-evolving residues (Lee et al, 2008; Suel et al, 2003; Lockless and Ranganathan, 1999; Shulman et al, 2004; Reynolds et al, 2011; Halabi et al, 2009), or local conservation in structure (Panjkovich and Daura, 2010). In related studies, both conservation and geometric-based searches for allosteric sites have been successfully applied to a few systems (Capra et al, 2009), several of which also employ support vector machines (Huang et al, 2006, Huang et al, 2013). Normal modes analysis, coupled with ligands of varying size, have been used to examine the extent to which bound ligands interfere with low-frequency motions, thereby identifying potentially important residues at the surface (Panjkovich and Daura, 2012; Mitternacht and Berezovsky, 2011; Ming and Wall, 2005).

In addition, the concept of 'protein quakes' has been introduced to explain local regions of proteins which are essential for conformation transitions (Miyashita et al

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** a consideration of

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** multiple conformations

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** an ensemble of structures

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** thereby rewiring essential pathways of information flow within the interior. Internal

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** SVMs

2003). A protein may relieve the strain of a high-energy configuration by local structural changes. Such local changes are often at the focal point of allosteric behavior, and these regions can be identified in a number of ways, including modified normal modes analysis (Miyashita et al 2003) or time-resolved X-ray scattering (Arnlund et al, 2014).

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

Though having provided valuable insights, many of these approaches may be limited in terms of scale (the numbers of proteins which may be feasibly investigated), computational demands, or the class of residues to which the method is tailored (surface or interior).

Using models of protein conformational change, we determine both surface and interior residues that may serve as essential allosteric regions in a computationally tractable manner, thereby enabling high-throughput analysis. This framework directly incorporates information regarding protein dynamics (as oppose to using less direct measures such as sequence features). In addition, this method is applied to a high-confidence set of proteins which exhibit conformational change. There is now a great deal of redundancy in folds and proteins, in that there are many proteins for which alternative crystal structures are available. This redundancy opens the door to large-scale analyses aimed at conformational heterogeneity and allosteric behavior on a database-level scale. We note that the residues identified tend to be conserved both across species and amongst humans. In a similar manner, several of our identified sites correspond to human disease loci for which no clear mechanism had previously been proposed for their pathogenicity. Finally, our pipeline (termed STRESS, for STRucturally-identified

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** communities-based

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** communities-based

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** essential

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** We describe a framework to identify instances of alternative conformations for a diverse set of proteins, and apply this to the PDB.

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** then

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** potentially

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** . In sum, we combine data from deep sequencing and cross-species conservation with the PDB database to identify potentially allosteric residues at varying protein depths in a manner that is both computationally efficient and directly aware of protein conformational changes and topology.

ESSential residues) is made available through a web server to which users may submit their own structures for analysis.

# RESULTS

## Predictions of Allosteric Residues

Allosteric residues at the surface generally play a regulatory role that is fundamentally different from that played by allosteric residues within the protein interior. While surface residues often represent the sources or sinks of allosteric signals (such as allosteric ligand binding sites in the former category, or distally located regulated sites in the latter), interior residues generally act to transmit allosteric signals across large distances. We use the models of protein conformational change as input for predicting both classes of allosteric residues. These predictions are first carried out using a gold standard set of 12 well-studied canonical systems for which both the *holo* and *apo* states are available (see below, Supp. Table 1, and Supp. Fig. 19), and are then implemented to predict allosteric residues on a database-level scale.

Models of conformational change may be taken directly from the vectors alternative crystal structures, or they may alternatively be inferred from anisotropic network models (ANMs), whereby the protein is modeled in a manner similar to that used in normal mode analysis. Here, interacting residues are modeled as nodes linked by flexible springs, in a manner similar to elastic network models and normal modes analysis. ANMs are simple and straightforward to apply on a database scale, and are thus used as our primary model of choice.

### *Identifying Critical Residues on the Surface*

We identify effector binding sites on the protein surface, some of which may act as latent ligand binding sites and active sites, using a modified version of the binding leverage framework for ligand binding site prediction (Mitternacht and Berezovsky, 2011, see Methods). Allosteric ligands often act by binding to surface cavities and modulate protein conformational dynamics. Thus, we first identify surface cavities using a series of Monte Carlo simulations to probe the protein surface with a simulated flexible

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** Identification

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** Potential

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** high-confidence set

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** alternative conformations described above

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** the

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** the

DECLAN CLARKE 8/16/15 1:26 PM
**Moved (insertion) [1]**

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** that

DECLAN CLARKE 8/16/15 1:26 PM
**Formatted:** Indent: First line: 0.5"

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** most important

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** allosteric regulation

DECLAN CLARKE 8/16/15 1:26 PM
**Moved (insertion) [2]**

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** Effector

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** are identified

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** modulating

ligand. The degree to which cavity occlusion by the simulated ligand disrupts large-scale conformational change is used to assign a score to each cavity (Fig. 1, bottom left; see Methods).

The main modifications to this formalism include the use of heavy atoms in the protein during the Monte Carlo search, in addition to an automated means of thresholding the list of ranked scores to give a more selective set of candidate sites. These modifications were implemented in order to provide a more selective set of residues: without them, we found that a very a large fraction of the protein surface would constitute predicted allosteric regions. We find that this modified approach results in finding an average of ~2 distinct binding sites per domain (Fig. 2a; see Methods for the details on defining distinct sites).

In order to evaluate the extent to which this method identifies known binding sites, we studied the ligand-binding sites within the gold standard set of proteins, and we positively identify an average of 60% of the sites known to be directly involved in ligand or substrate binding. Some of the sites identified do not directly overlap sites of known biological significance. However, such sites may nevertheless correspond to latent allosteric regions (Bowman et al, 2015): even if no known biological function is assigned to such sites, their occlusion may still disrupt large-scale motions. Secondly, we often find that these sites nevertheless exhibit some degree of overlap with sites of biological interest, suggesting that the identified sites often lie within the neighborhood of known biological sites (Supp. Table 4).

***Dynamical Network Analysis to Identify Critical Residues Within the Interior***

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

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** , as defined by either anisotropic network models (ANMs) or direct knowledge of alternative crystal structures (see below),

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:**

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** We find

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** this

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** and active sites in a

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** 12 well-studied systems for which the crystal structures of both the *holo* and *apo* states are available. We find that, out of the 12 canonical systems studied,

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** .

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** meet the thresholds needed for defining a site

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:**

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** described above

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** Thus, we apply communities-based network analyses to the proteins of our dataset to identify important internal residues. Such

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** as

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** surface sites. Modeling the protein structure as a network of interacting

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** , an edge between a given pair

available routes for information flow, thereby making these residues essential information flow bottlenecks.

To identify these bottlenecks, the protein may first be modeled as a network of interconnecting residues, 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 ANMs, see above). In this regard, the problem of identifying internal allosteric residues is reduced to a problem of identifying which nodes participate in network bottlenecks.

These bottlenecks are identified using an approach schematized in Supp. Fig. 16 (see Methods for details). Briefly, the network edges are first weighted by the correlated motions of contacting residues: a strong correlation in the motion between residues suggests 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. Using the 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; threonine synthase, for example, exhibits the community partition shown in Supp. Fig. 6. Finally, the betweenness of each edge (defined to be the number of shortest paths between all pairs of residues which pass through a that edge) is calculated, and those residues that are involved in the highest-betweenness interactions between all pairs of interacting communities are predicted to be allosteric. These residues are critical for information flow between communities, as their removal would result in substantially longer paths between the residues of one community to those of another, thereby substantially reducing the efficiency of information flow between communities.

USING EDGE WC+

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** designates a mutual proximity of 4.5 Angstroms. Edges

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** on the basis of

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** movements between

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** (see Methods).

DECLAN CLARKE 8/16/15 1:26 PM
**Formatted:** Font:(Default) Times New Roman, 12 pt, Not Bold, Not Italic

***Web-Based Tool: STRESS (STRucturally-identified ESSential residues)***

Both the surface- and interior-critical residue identification modules have been made available to the community through a new web server, STRESS. A user may specify a PDB to be analyzed, and the output provided constitutes the set of predicted allosteric residues. Obviating the need for long wait times, the algorithmic

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** ***Server***

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** Our server has been designed

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** both user-friendly

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** fast.

implementation of our software is highly efficient (running times for proteins of various sizes are provided in Fig. 5). In the surface-critical residue identification module, we use local searching to bring down the time complexity by an order comparing with a naïve implementation. After carefully profiling and optimization, a typical case takes only about 30 minutes on one typical CPU (2.8GHz) core.

Running time is also minimized by designing a scalable server architecture. The thin front-facing servers handle incoming user requests, and more powerful back-end servers perform calculations. The back-end servers are automatically and dynamically scalable, ensuring that it can handle varying levels of demand. This implementation is based on Amazon Web Service (AWS) and is highly portable on cloud environment.

# Models of Protein Conformational Change

### *High-Throughput Identification of Structures in Distinct Energetic Wells*

Protein conformational change is an principal component and assumption in our identification of important residues. Thus, to better ensure that the proteins studied exhibit well-characterized distinct conformations, we use a generalized approach to systematically identify instances of proteins that exhibit alternative conformations within the PDB.

As a first step, we perform multiple structure alignments (MSAs) across sequence-identical domains as well as proteins, with these structures having been filtered by resolution and other metrics to ensure quality (Fig. 1). We then use the resultant pairwise RMSD values to infer distinct conformational states (see Methods). The distribution of the resultant number of conformations for domains and chains is given in Fig. 2D and 2E, respectively. Results remained the same whether we used RMSD or $Q_H$ to quantify similarity (Supp. Fig. 3), and we use RMSD in our downstream analyses. The fully-processed output of alternative conformations is provided in Supp. File 1, and the conformational transitions we observe arise in a diverse set of biological contexts, including conformational changes that accompany ligand binding, protein-protein or protein-nucleic acid interactions, post-translational modifications, and changes in oxidation or oligomerization state. (Supp. Fig. 4).

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** locality-sensitive hashing to do

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** search in each sampling step, which takes constant time. The

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** of the core computation, Monte Carlo sampling, is O(|T||S|), where T and S are simulation trials and steps for each trial, respectively.

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** E5-2650

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** **([[STL2MG]]need to confirm with Mihali/Mark, what kind of core we purchased on Grace)** core.

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** In terms of server operation, our web application utilizes two types of servers: front-facing servers that handle incoming HTTP requests and back-end servers that perform algorithmic calculations. Communication between these two types of servers is handled by Amazon's Simple Queue Service. When our front-facing servers receive a new request, they add the job to the queue and then return to handling requests immediately. Our back-end servers continually poll the queue for new jobs and run them when capacity is available. Amazon's Elastic Beanstalk offers several features that enable us to dynamically scale our web application. We use Auto Scaling to automatically adjust the number of servers backing our application based on predefined conditions, such as network traffic and CPU utilization. Elastic Load Balancer then automatically distributes incoming traffic across these servers. This system ensures that we are able to handle varying levels of demand in a reliable and cost-effective manner. Since we may have multiple servers backing our web application ... [1]

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** Knowledge of protein

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** heterogeneity and

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** essential

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** essential

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** in

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** throughout

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:**

An overview of our dataset in the broader context of the entire PDB is given in Supp. Fig. 17. In addition, the distribution of the number of chains and domains for our dataset are given in Supp. Fig. 18a and b, respectively.

RESOURCE

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

As mentioned, directly using the displacement vectors between all corresponding pairs of residues within the two crystal structures of the alternative conformations provides another model of conformational change, and we find that this alternative gives the same general results (see Supp. Fig. 15 and Supplemental discussion). Thus, our method is general with respect to how motion vectors are defined.

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** Models of conformational change may be taken directly from the vectors describing alternative crystal structures, or they may alternatively be inferred from ANMs, whereby the protein is modeled in a manner similar to that used in normal mode analysis.

DECLAN CLARKE 8/16/15 1:26 PM
**Moved up [2]:** Here, interacting residues are modeled as nodes linked by flexible springs, in a manner similar to elastic network models and normal modes analysis. ANMs are simple and straightforward to apply on a database scale, and are thus used as our primary model of choice

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** .

## Conservation Analyses on Critical Residues

Applying the efficient allosteric site prediction formalism to the large number of dynamic proteins culled throughout the PDB enables a large-scale analysis of the residues identified. In particular, and obvious question that arises is the extent to which the sites identified tend to be conserved. Thus, we evaluate the conservation of the predicted allosteric sites for this large set of proteins, with conservation being evaluated both across long (inter-species) and short (intra-human) evolutionary timescales. 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 such conservation also reinforces the functional importance of the predicted allosteric sites.

***Conservation Across Species***

Predicted allosteric residues tend to be more conserved, on average, than other residues of the same protein with the same degree of burial, and these results hold for both surface- and interior-critical residues (Figs. 3B and 3F, respectively). Surface critical residues had an average ConSurf score ("conservation score") of -0.131, whereas non-critical residues with the same degree distribution (i.e., same degree of burial within the protein) had an average score of +0.059, demonstrating that surface-critical residues tend to be more conserved ($p < 2.2e-16$). Interior-critical residues exhibit a similar trend: the

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** Our identified critical

average conservation score for interior critical residues and non-critical residues is -0.179 and -0.102, respectively (p=3.67e-11).

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

Though inter-species metrics may be used to investigate conservation, we may also use the genomes and exomes from thousands of humans to study selective constraints which are both human-specific and active in more recent evolutionary history. Common metrics for such constrains include allele rarity and the fraction of rare alleles (cite GOSB review?).

Although we observe a general trend in which rare alleles from 1000 Genomes coincide with surface critical residues, the trend is not observed to be significant (Fig. 3C; p=0.309). The significance improves when considering the shift in the allele frequencies, as evaluated with a K-S test (p=0.08, Supp. Fig. 13a), and we note the limited number of proteins (44) to be hit by 1000 Genomes single-nulceotide variants (SNVs; see Methods). The long tail extending to lower allele frequencies for critical residues may suggest the possibility that only a subset of residues in our prioritized binding sites is essential. Notably, 1000 Genomes variants hit critical-interior residues with significantly lower derived allele frequency than non-critical residues with the same degree (Fig. 3G).

We also performed a similar analysis using the data provided by the Exome Aggregation Consortium (Exome Aggregation Consortium, abbreviated ExAC). The trends obtained using ExAC are similar to those using 1000 Genomes data (distributions for critical-surface and critical-interior residues are given in Figs. 3D and 3H, respectively). Although the mean minor allele frequencies (MAF) for surface-critical residues are higher than those of non-critical residues, the median for surface-critical residues is substantially lower than that for non-critical residues. The relative shifts of these distributions are also shown in Supp. Fig. 14 (KS test p= 0.0475 and p= 8.7E-5 for critical-surface and critical-interior residues, respectively).

In addition to examining allele frequency distributions, one may also evaluate the *fraction* of rare alleles as a metric for measuring selective pressure (defined as the ratio of the number of low-DAF or low-MAF SNVs to all non-synonymous SNVs in a given protein). Using different DAF cutoffs for 1000 Genomes variants (0.5% and 0.1%) to

define rarity, the results for surface- and interior-critical residues are summarized in Supp. Fig. 7 and Supp. Fig. 8, respectively. Similar results are obtained when using ExAC variants: we find that surface residues are generally more conserved than other residues, and this result holds using different thresholds for defining rarity (Supp. Table 7). In sum, when using different thresholds for defining rarity, critical residues tend to be enriched in rare variants, again suggesting their greater degree of selection.

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

Directly related to conservation is the concept of variant deleteriousness: changes in amino acid composition at specific loci may be more or less likely to result in disease. SIFT and PolyPhen are two tools for predicting such effects, and we evaluated these predictions for critical and non-critical residues hit by variants in ExAC. Variants hitting critical residues exhibit significantly higher PolyPhen scores relative to non-critical residues, suggesting the potentially higher disease susceptibility at critical residues (Supp. Fig. 12; higher PolyPhen scores denote more damaging variants), though significant disparities were not observed in SIFT scores (Supp. Fig. 11).

Using HGMD, we identify several proteins to be hit by known disease mutations, (Fig. 4A and Supp. Files 2 – 5; Stenson et al 2014). Several identified critical residues coincide with known disease loci for which the mechanism of pathogenicity is unclear unless an allosteric cause is considered.

Fibroblast growth factor receptor is a case-in-point (Fig. 2F and Supp. Table 6), variants in which have been linked to diseases that manifest in craniofacial defects. Dotted lines highlight poorly-understood disease variants that coincide with our critical residues. The incorporation of critical surface and interior residues introduces an additional layer of annotation to the protein sequence, and may thus help to explain otherwise poorly understood disease variants.

DECLAN CLARKE 8/16/15 1:26 PM
**Moved up [1]:** Fig.

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** Though no significant disparity was observed in SIFT scores (Supp.

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** 11), variants

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** -

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** or critical-

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** adds

# DISCUSSION & CONCLUSIONS

The same principles of energy landscape theory that dictate protein folding are integral to how proteins explore different conformations once they adopt their folded

states. These landscapes are shaped not only by the protein sequence itself, but also by extrinsic conditions. Such external factors often regulate protein activity by introducing allosteric-induced changes, which ultimately reflect changes in the shapes and population distributions of the energetic landscape. In this regard, allostery provides an ideal platform from which to study protein behavior in the context of their energetic landscapes.

To investigate allosteric regulation, and to simultaneously add an extra layer of annotation to each protein in the context of its conservation patterns, an integrated framework to identify allosteric residues throughout the protein is essential. We introduce a framework to identify essential residues that leverages knowledge of conformational heterogeneity. To identify potential allosteric residues closer to the surface, heavy atoms are included when searching the surface for sites in which the introduction of a ligand could strongly perturb conformational changes. Secondly, after these sites are identified, we use a formalism originally used in the context of protein folding (the energy gap [[cite]]), to define a threshold for selecting the high-confidence prioritized sites. The set of high-confidence sites overlaps reasonably well with known ligand binding sites for several well-studied canonical allosteric systems.

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

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

Unlike the characterization of many other structural features, such as secondary structure assignment, residue burial, protein-protein interaction interfaces, disorder, and

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** .

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** Though a small number of examples in which allostery can occur without conformational change have been discussed in the literature (Tsai et al, 2009; Nussinov et al, 2015), the fact that these specific systems have been highlighted as exceptions underscores the ubiquity of allosteric regulation.

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** This ubiquity suggests that, not only is allostery an essential component to understanding protein behavior in general, but it is also well suited to provide a conceptual framework for understanding many of the conserved regions found in proteins. Such conservation patterns are increasingly coming to light in the age of next-generation sequencing. As such, higher-throughput approaches for identifying potentially allosteric residues are needed to meet the high- ... [2]

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** potentially

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** Notably, various methods fo ... [3]

DECLAN CLARKE 8/16/15 1:26 PM
**Moved down [3]:** MD and NMR ar ... [4]

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** Given the limitations in appl ... [5]

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** hybrid method

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** we describe a framework in which

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** search for

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** modules

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** We apply this framework

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** a large number of the alterna ... [6]

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** for which alternative crystal ... [7]

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** heterogeneity and potential ... [8]

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** our critical

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** . The critical residues identif ... [9]

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** HGMD was used in order t ... [10]

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

There ~~are~~ several notable implications of our database-scale analysis. Static protein feature (such as those listed above) are generally much more accessible than dynamic features, whereas allostery is 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 allosteric hotspots in protein-protein interaction networks. Knowledge of predicted allosteric sites across many proteins may also be used to identify the best proteins and protein regions for which drugs should be engineered, as well as instances in which specific sequence variants are likely to have the greatest impact.

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

The ability to leverage our framework in a high-throughput also better enables one to match structural features with the high-throughput data generated through deep sequencing. Full human genomes and exomes are being sequenced at an increasing pace, thereby providing an unprecedented window into conservation patterns which can be

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** Given that

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** has previously been studied in the context of individual proteins, there

DECLAN CLARKE 8/16/15 1:26 PM
**Moved (insertion) [3]**

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** en masse suggests avenues for investigating

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** or guiding experimental studies to prioritize residues that are candidates for allosteric behavior (cite Rama Ranganathan, others).

human-specific or active over short evolutionary timescales. With such large volumes of data, these patterns increasingly serve as detailed signatures, or as it were, "shadows" 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.

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

We also anticipate that our newly-developed server (STRESS) will prove to be useful in these and related studies (<URL_HERE>). Users may submit protein structures in order to perform their own analyses for predicting allosteric residues. As next-generation sequencing initiatives continue to provide a clearer picture of conservation at the residue level, structural biologists will increasingly find a need to explain the emergent conservation patterns, and that our server readily enables the search for allosteric regions.

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** identifying essential

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** potential

# METHODS

DECLAN CLARKE 8/16/15 1:26 PM
Deleted:

An overview of our pipeline is provided in Fig. 1, and we refer to this outline in the appropriate pipeline modules throughout. In brief, we perform MSAs for thousands of SCOP domains, with each alignment consisting of sequence-similar and sequence-identical domains. Within each alignment, we cluster the domains using structural similarity to determine the distinct conformational states. We then implement coarse-grained models of protein motions to identify allosteric sites on the protein surface, as well as dynamical network analysis to identify allosteric residues internal to the protein.

### Database-Wide Multiple Structure Alignments

FASTA files of all SCOP domains were downloaded from the SCOP website (version 2.03) [[cite]]. In order to better ensure that large structural differences between sequence-identical or sequence-similar domains are a result of differing biological states (such as *holo* vs. *apo*, phosphorylated vs. unphosphorylated, etc.), and not an artifact of missing coordinates in X-ray crystal structures, the FASTA sequences used were those corresponding to the ATOM records of their respective PDBs. In total, this set comprises 162,517 FASTA sequences.

BLASTClust [[cite]] was downloaded from the NCBI database and used to organize these FASTA sequences into sequence-similar groups at seven levels of sequence identity (100%, 95%, 90%, 70%, 50%, 40%, and 30%). Thus, for instance, running BLASTClust with a parameter value of 100 provides a list of FASTA sequence groups such that each sequence within each group is 100% sequence identical, and in general, running BLASTClust with any given parameter value provides sequence groups such that each member within a group shares at least that specified degree of sequence identity with any other member of the same group (see top of Fig. 1).

To ensure that the X-Ray structures used in our downstream analysis are of sufficiently high quality, we removed all of those domains corresponding to PDB files with resolution values poorer than 2.8, as well as any PDB files with R-Free values poorer than 0.28. The question of how to set these quality thresholds is an important consideration, and was guided here by a combination of the thresholds conventionally

used in other studies which rely on large datasets of structures [[cite Kosloff 2008, Burra 2009, others]], as well as the consideration that many interesting allosteric-related conformational changes may correlate with physical properties that sometimes render very high resolution values difficult (such as localized disorder or order-disorder transitions). As a result of applying these filters, 45,937 PDB IDs out of a total of 58,308 unique X-Ray structures (~79%) were kept for downstream analysis.

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

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

Finally, for each of the N MSAs generated, MultiSeq was used calculate two measures of structural similarity between each pair of domains within a group: RMSD and $Q_H$. A fuller description of $Q_H$ is provided in the Supplementary text.

For each group of sequence-similar domains, the final output of the structure alignment is a symmetric matrix representing all pairwise RMSD values (as well as a separate matrix representing all pairwise $Q_H$ values) within that group. The matrices for all MSAs are then used as input to the K-means module.

### Identifying Distinct Conformations in an Ensemble of Structures

For each MSA produced in the previous step, the corresponding matrix of pairwise RMSD values describes the degree and nature of structural heterogeneity among the crystal structures for a particular domain. The objective is to use this data in order to identify the biologically distinct conformations represented by an ensemble of structures.

For a particular domain, there may be many available crystal structures. In total, these structures may actually represent only a small number of distinct biological states and conformations. For instance, there may be several crystal structures in which the domain is bound to its cognate ligand, while the remaining structures are in the *apo* state. Our framework for predicting the number of distinct conformational states in an ensemble of structures relies on a modified version of the K-means clustering algorithm.

A priori, performing K-means clustering assumes prior knowledge of the number of clusters (i.e., "K") to describe a dataset. The purpose of K-means clustering with the gap statistic (Tibshirani et al, 2001) is to identify the optimal number of clusters intrinsic to a complex or noisy set of data points (which lie in N-dimensional space).

Given multiple resolved crystal structures for a given domain, this method (i.e., K-means with the gap statistic) estimates the number of conformational states represented in the ensemble of crystal structures (with these states presumably occupying different wells within the energetic landscape), thereby identifying proteins which are likely to undergo conformational change as part of their allosteric behavior.

As a first step toward clustering the structure ensemble represented by the RMSD matrix, it is necessary to convert this RMSD matrix (which explicitly represents only the *relationships* between distinct domains) into a form in which each domain is given its own set of coordinates. This step is necessary because the K-means algorithm acts directly on individual data points, rather than the distances between such points. Thus, we use multidimensional scaling [[ref Gower 1966 and Mardia, 1978]] to convert an N-by-N matrix (which provides all RMSD values between each pair of domains within a group of N structures) into a set of N points, with each point representing a domain in (N-1)-dimensional space. The values of the N-1 coordinates assigned to each of these N points are such that the Euclidean distance between each pair of points are the same as the RMSD values in the original matrix. For an intuition into why N points must be mapped to (N-1)-dimensional space, consider an MSA between two structures. The RMSD between these two structures can be used to map the two domains to one-dimensional space, such that the distance between the points is the RMSD value. Similarly, an MSA of 3 domains may be mapped to 2-dimensional space in such a way that the pairwise distances are preserved; 4 domains may be mapped to 3-dimensional space, etc. The

output of this multidimensional scaling is used as input to the K-means clustering with the gap statistic. We refer the reader to the work by Tibshirani et al for details governing how we perform K-means clustering with the gap statistic.

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

To validate the output generated by this clustering algorithm, we manually annotated the alignments of a vast array well-studied canonical allosteric domains and proteins. There may be many factors driving conformational change, and those cases for which the change is induced by the binding to a simple ligand (i.e., a consideration of *apo* or *holo* states) constitute only a very small subset of the conformational shifts observed in the PDB. For instance, such shifts often result from protein-protein or protein-nucleic acid interactions, changes in oxidation states or in pH, mutations, binding to very large and complex ligands or the potential to bind to variable sets of ligands, post-translational modifications, interactions with the membrane, shifts in oligomerization states or configuration, etc. The gap statistic performed well in discriminating crystal structures that constitute such a diverse set, and this method has been validated using both domains (Supp. Figs. 4a-f) and protein chains (Supp. Figs. 4g-x).

RMSD values were used to generate dendrograms for each of the selected MSAs. The dendrograms are constructed using the hierarchical clustering algorithm built into R, `hclust` [[ref Murtagh 1985]], with UPGMA (mean values) used as the chosen agglomeration method[[ref Sokal et al, 1958]].

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

We then select a representative domain from each of the assigned clusters. The representative member for each cluster is the member with the lowest Euclidean distance to the cluster mean, using the coordinates obtained by multidimensional scaling (see

description above). These cluster representatives are then taken as the distinct conformations for this protein, and are used for the binding leverage calculations and networks analyses (below).

### Modified Binding Leverage Framework

With the objective of identifying allosteric residues (specifically those on the protein surface), we employed a modified version of the binding leverage method for predicting likely ligand binding sites (Fig. 1, bottom-left), as described previously by Mitternacht and Berezovsky. Allosteric signals may be transmitted over large distances by a mechanism in which the allosteric ligand has a global affect on a protein's functionally important motions. For instance, introducing a bulky ligand into the site of an open pocket may disrupt large-scale motions if those motions normally entail that the pocket become completely collapsed in the *apo* protein. Such a modulation of the global motions may affect activity within sites that are distant from the allosteric ligand-binding site.

We refer the reader to the work by Mitternacht and Berezovsky for details regarding the binding leverage method, though a general overview of the approach follows. Many candidate allosteric sites are generated by simulations in which a simple ligand (comprising 2 to 8 atoms linked by bonds with fixed lengths but variable bond and dihedral angles) explores the protein's surface through many Monte Carlo steps (*apo* structures were used when probing protein surfaces for putative ligand binding sites). A simple square well potential (i.e., modeling hard-sphere interactions) was used to model the attractive and repulsive energy terms associated with the ligand's interaction with the surface. These energy terms depend only on the ligand atoms' distance to alpha carbon atoms in the protein, and they are blind to other heavy atoms or biophysical properties. Once these candidate sites have been produced, normal mode analysis is applied is generate a model of the *apo* protein's low-frequency motions. Each of the candidate sites is then scored based on the degree to which deformations in the site couple to the low-frequency modes; that is, those sites which are heavily deformed as a result of the normal mode fluctuations receive a high score (termed the binding leverage for that site), whereas sites which undergo minimal change over the course of a mode fluctuation

receive a low binding leverage score. The list of candidate sites is then processed to remove redundancy, and then ranked based on this score. The model stipulates that the high-scoring sites are those that are more likely to be binding sites. Using knowledge of the experimentally-determined binding sites (i.e., from *holo* structures), the processed list of ranked sites is then used to evaluate predictive performance (see below).

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

In the original binding leverage framework, an interaction between a ligand atom and an alpha carbon atom in the protein contributes -0.75 to the binding energy if the interaction distance is within the range of 5.5 to 8 Angstroms. Interaction distances greater than 8 Angstroms do not contribute to the binding energy, but distances in the range of 5.0 to 5.5 are repulsive, and those between 4.5 to 5.0 Angstroms are strongly repulsive (distances below 4.5 Angstroms are not permitted).

DECLAN CLARKE 8/16/15 1:26 PM
**Deleted:** origi

However, given the much higher density of atoms interacting with the ligand in our all-heavy atom model of each protein, it is necessary to accordingly change the energy parameters associated with the ligand's binding affinity. In particular, we varied both the ranges of favorable and unfavorable interactions, as well as the attractive and repulsive energies themselves (that is, we varied both the square well's width and depth when evaluating the ligand's affinity for a given site).

For well depths, we employed models using attractive potentials ranging from -0.05 to -0.75, including all intermediate factors of 0.05. For well widths, we tried performing the ligand simulations using the cutoff distances originally used (attractive in the range of 5.5 to 8.0 Angstroms, repulsive in the range of 5.0 to 5.5, and strongly

repulsive in the range of 4.5 to 5.0). However, these cutoffs, which were originally devised to model the ligand's affinity to the alpha carbon atom skeleton alone, were observed to be inappropriate when including all heavy atoms. Thus, we also performed the simulations using a revised set of cutoffs, with attractive interactions in the range of 3.5 to 4.5 Angstroms, repulsive interactions in the range of 3.0 to 3.5 Angstroms, and strongly repulsive interactions in the range of 2.5 to 3.0 Angstroms.

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

The biological assembly files were downloaded from the Protein Data Bank (PDB). These proteins were chosen on the basis of literature curation.

### Network Analysis

In our implementation of the Girvan-Newman framework, edges between residues within a structure are drawn between any two residues that have at least one heavy atom within a distance of 4.5 Angstroms (excluding adjacent residues in sequence, which are not considered to be in contact). Network edges are weighted on the basis of their correlated motions, with the motions provided by ANMs. We emphasize that, although the use of ANMs is more coarse-grained that MD, our use of ANMs is motivated by their much faster computational efficiency. This added efficiency is a required feature for our database-scale analysis.

Specifically, the weight $w_{ij}$ between residues i and j is set to $-\log(|C_{ij}|)$, where $C_{ij}$ designates the correlated motions between residue i and j. If two contacting residues exhibit a high degree of correlated motion, then this implies that the motion of one residue may tell us about the motion of the other, suggesting a strong flow of energy or information between the two residues, resulting in a low value for $w_{ij}$. The 'network distance' between residues i and j (synonymous with $w_{ij}$ in this discussion) is thus taken

to be very short, and this short distance means that any path involving this pair of residues is shorter as a result, thereby more likely placing this pair of residues within any given shortest path, and more likely rendering this pair of residues a bottleneck pair. In sum, a high correlation in motion results in a short distance, thereby more likely rendering this a bottleneck pair of residues.

Finally, once all connections between contacting pairs are appropriately weighted and the communities are assigned, a residue is deemed to be critical for allosteric signal transmission if it is involved in a highest-betweenness edge connecting two distinct communities. For instance, applying this method to threonine synthase results in the community partition and associated critical residues highlighted in Supp. Fig. 6.

### Conservation Analyses

All cross-species conservation scores represent the ConSurf scores, as taken from the ConSurf Server [[cite]], in which scores for each protein chain are normalized to 0. Low (negative) ConSurf scores represent a stronger degree of conservation, and high (positive) scores designate less stringent selection. Each point within the cross-species conservation plots (Figs 3B and 3F) represents the mean conservation score for all residues within one of two classes: the full set of N critical residues within a protein structure or a randomly-selected set of N non-critical residues (with the same degree) within the same structure. The randomly-selected non-critical set of residues was chosen in a way such that, for each critical residue with degree K (K being the number of non-adjacent residues with which the critical residue is in contact), a randomly-chosen non-critical residue with the same degree K was included in the set. The distribution of non-critical residues shown is very much representative of the distribution observed when re-building the random set many times.

Our use of degree as a metric for characterizing burial is consistent with our networks-based analysis for identifying interior critical residues, as well as our use of residue-residue contacts in building networks for producing the ANMs. Residue degree is also an attractive metric because it is discrete in nature, thereby allowing us to generate null distributions of non-critical residues with the exact same degree distribution.

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

ExAC variants were downloaded from the ExAC Browser (Beta), as hosted at the Broad Institute. Variants were mapped to all PDBs following the same protocol as that used to map 1000G variants, and only non-synonymous SNVs in ExAC were analyzed. When evaluating SNVs from the ExAC dataset, minor allele frequencies were used instead of DAF values (the ancestral allele is not provided in the ExAC dataset – thus, analysis is performed for MAF rather than DAF. However, we note that very little difference was observed when using AF or DAF values with 1000G data, and we believe that the results with MAF values would generally be the same to those with DAF values). Only structures for which at least one critical residue and one non-critical residue are hit by ExAC SNVs are included in the analysis (as with the 1000 Genomes analysis, this enables a more direct comparison between critical and non-critical residues, as comparisons between two different proteins would rely on the assumption of equal degrees of selection between such proteins).

## Web Server (STRESS)

Our server has been designed to be both user-friendly and fast. As discussed, we use locality-sensitive hashing to do local search in each sampling step in the search for surface-critical residues, which takes constant time. The time complexity of the core computation, Monte Carlo sampling, is O(|T||S|), where T and S are simulation trials and steps for each trial, respectively. After carefully profiling and optimization, a typical case takes only about 30 minutes on one E5-2650(2.8GHz) **([[STL2MG]]need to confirm with Mihali/Mark, what kind of core we purchased on Grace)** core.

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

CODE IGLT

# FIGURE CAPTIONS

**Figure 1**

**Pipeline for identifying distinct conformational states.** *Top to bottom*: **a)** BLAST-CLUST is applied to the sequences corresponding to a filtered set of protein domains, thereby providing a large number of "sequence groups", with each group being

characterized by a high degree of sequence homology. **b)** For each sequence group, a multiple structure alignment of the domains is performed using STAMP (the example shown here is adenylate kinase. The SCOP IDs of the cyan domains, which constitute the *holo* structure, are d3hpqb1, d3hpqa1, d2eckb1, d2ecka1, d1akeb1, and d1akea1. The IDs of the *apo* domains, in red, are d4akea1 and d4akeb1). **c)** Using the pairwise RMSD values in this structure alignment, the structures are clustered using the UPGMA algorithm, K-means with the gap statistic (δ) is performed to identify the number of distinct conformations (2 in this example; more detailed descriptions of the graph are provided in the text). **d)** The domains which exhibit multiple structural clusters (i.e., those with a $\delta > X$ and $K > 1$) are then probed for the presence of strong allosteric sites, using binding leverage and dynamical network analysis (see Methods).

**Figure 2**

**K-means clustering algorithm with the gap statistic.** Number of binding sites per domain **(a)** and complex **(b)**; **c)** An example dendrogram and respective structures of a multiple-structure alignment, with similarity measured by RMSD. The example shown is for phosphotransferase, and the K-means algorithm with the gap statistic identifies K=2 different conformational states (manually determined to represent the *holo* and *apo* states of phosphotransferase); **d)**Histograms representing the K-values obtained across the database of SCOP domains and **e)** across PDB chains. Shown in **(f)** is a linear annotation diagram for fibroblast growth factor receptor. Shown is chain E of the PDB 1IIL, which corresponds to the FGFR2. Dotted lines highlight loci that correspond to HGMD sites that coincide with critical residues, but for which other annotations fail to coincide. Deeply-buried residues are defined to be those that exhibit a relative solvent-exposed surface area of 5% or less, and binding site residues are defined as those for which at least one heavy atom falls within 4.5 Angstroms of any heavy atom in the binding partner (heparin-binding growth factor 2). The loci of PTM sites were taken from UniProt (accession no. P21802).

**Figure 3**

**Conservation of predicted allosteric residues.**

Throughout, red designates critical residues, and blue designates non-critical residues, and results are reported for all proteins in our database with available ConSurf scores (cross-species plots) and all proteins hit by a variant in at least one critical and one non-critical residue (1000 Genomes and ExAC plots). P values are calculated using a Wilcoxon Rank sum test. **a)** Image of phosphfructokinase (PDB ID 3PFK), with red denoting sites with high binding leverage scores, and blue denoting sites with low scores; **b)** Distributions of mean conservation scores for surface-critical and non-critical residues ($p < 2.2e-16$); **c)** Distributions of mean derived allele frequencies (DAF) of 1000 Genomes variants on surface-critical and non-critical residues ($p=0.309$); **d)** Distributions of mean minor allele frequencies (MAF) of ExAC variants on critical-surface and non-critical residues ($p=1.49e-3$); **e)** Rendering of phosphfructokinase with interior critical residues highlighted as red spheres; **f)** Distributions of conservation scores for interior-critical residues and non-critical residues ($p=9.31e-11$); **g)** Distributions of DAF values for 1000 Genomes variants hitting interior-critical residues and non-critical residues ($p=1.80e-05$); **h)** Distributions of mean MAF values for ExAC variants hitting critical-interior residues and non-critical residues ($p=7.98e-09$).

**Figure 4**

**HGMD Analyses. a)** Venn diagram illustrating the number of distinct proteins in various categories; **b)** Ras (PDB ID 1NVV) is an example of a protein for which HGMD locations coincide with prioritized sites. Surface critical residues are shown as red spheres, and HGMD locations are in orange; **c)** p53 (PDB ID 2VUK) is an example of a protein for which HGMD locations coincide with interior critical residues. Interior critical residues that coincide with HGMD SNVs (red), critical residues that do not correspond with HGMD loci (green), and HGMD SNVs in non-critical residues (orange) are shown in VDW spheres.

# REFERENCES

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

Arora, Karunesh, and Charles L. Brooks. "Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism." Proceedings of the National Academy of Sciences 104.47 (2007): 18496-18501.

Ashkenazy, Haim, et al. "ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids." Nucleic acids research (2010): gkq399.

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

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

Bowman, Gregory R., et al. "Discovery of multiple hidden allosteric sites by combining Markov state models and experiments." Proceedings of the National Academy of Sciences 112.9 (2015): 2734-2739.

Burra, Prasad V., et al. "Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure." Proceedings of the National Academy of Sciences 106.26 (2009): 10505-10510.

Capra, John A., et al. "Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure." PLoS Comput Biol 5.12 (2009): e1000585.

Celniker, Gershon, et al. "ConSurf: using evolutionary data to raise testable hypotheses about protein function." Israel Journal of Chemistry 53.3‐4 (2013): 199-206.

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

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

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

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

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

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

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

Fox, Naomi K., Steven E. Brenner, and John-Marc Chandonia. "SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures." Nucleic acids research 42.D1 (2014): D304-D309.

Gasper, Paul M., et al. "Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities." Proceedings of the National Academy of Sciences 109.52 (2012): 21216-21222.

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

Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." Proceedings of the National Academy of Sciences 99.12 (2002): 7821-7826.

Glaser, Fabian, et al. "ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information." Bioinformatics 19.1 (2003): 163-164.

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

Gower, J. C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53, 325–328.

Grant, Gregory A., David J. Schuller, and Leonard J. Banaszak. "A model for the regulation of D‐3‐phosphoglycerate dehydrogenase, a Vmax‐type allosteric enzyme." Protein science 5.1 (1996): 34-41.

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

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

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

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

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

Kohl, Andreas, et al. "Allosteric inhibition of aminoglycoside phosphotransferase by a designed ankyrin repeat protein." Structure 13.8 (2005): 1131-1141

Kosloff, Mickey, and Rachel Kolodny. "Sequence‐similar, structure‐dissimilar protein pairs in the PDB." Proteins: Structure, Function, and Bioinformatics 71.2 (2008): 891-902.

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

Lancichinetti, Andrea, and Santo Fortunato. "Community detection algorithms: a comparative analysis." Physical review E 80.5 (2009): 056117.

Landau, Meytal, et al. "ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures." Nucleic acids research 33.suppl 2 (2005): W299-W302.

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

Laurent, M., et al. "Solution X-ray scattering studies of the yeast phosphofructokinase allosteric transition. Characterization of an ATP-induced conformation distinct in quaternary structure from the R and T states of the enzyme." Journal of Biological Chemistry 259.5 (1984): 3124-3126.

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

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

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

Manley, Gregory, Ivan Rivalta, and J. Patrick Loria. "Solution NMR and computational methods for understanding protein allostery." The Journal of Physical Chemistry B 117.11 (2013): 3063-3073.

Mardia, K.V. (1978) Some properties of classical multidimensional scaling. Communications on Statistics – Theory and Methods, A7, 1233–41.

Ming D, Wall ME: Quantifying allosteric effects in proteins. Proteins 2005, 59(4):697-707.

Mitternacht, Simon, and Igor N. Berezovsky. "Binding leverage as a molecular basis for allosteric regulation." PLoS computational biology 7.9 (2011): e1002148.

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

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

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

Panjkovich, Alejandro, and Xavier Daura. "Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery." BMC structural biology 10.1 (2010): 9.

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

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

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

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

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

N Tibshirani, Robert, Guenther Walther, and Trevor Hastie. "Estimating the number of clusters in a data set via the gap statistic." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63.2 (2001): 411-423.

Tsai, Chung-Jung, Buyong Ma, and Ruth Nussinov. "Folding and binding cascades: shifts in energy landscapes." Proceedings of the National Academy of Sciences 96.18 (1999): 9970-9972.

Tsai, Chung-Jung, Antonio Del Sol, and Ruth Nussinov. "Allostery: absence of a change in shape does not imply that allostery is not at play." Journal of molecular biology 378.1 (2008): 1-11.

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

Rosvall, Martin, and Carl T. Bergstrom. "An information-theoretic framework for resolving community structure in complex networks." Proceedings of the National Academy of Sciences 104.18 (2007): 7327-7331.

Sethi, Anurag, et al. "Dynamical networks in tRNA: protein complexes." Proceedings of the National Academy of Sciences 106.16 (2009): 6620-6625.

Sethi, Anurag, et al. "A mechanistic understanding of allosteric immune escape pathways in the HIV-1 envelope glycoprotein." PLoS computational biology 9.5 (2013): e1003046.

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

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

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

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

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

Wiesmann, Christian, et al. "Allosteric inhibition of protein tyrosine phosphatase 1B." Nature structural & molecular biology 11.8 (2004): 730-737.

Xiang, Yun, et al. "Simulating the effect of DNA polymerase mutations on transition-state energetics and fidelity: Evaluating amino acid group contribution and allosteric coupling for ionized residues in human pol β." Biochemistry 45.23 (2006): 7036-7048.

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

VanWart, Adam T., et al. "Exploring residue component contributions to dynamical network models of allostery." Journal of chemical theory and computation 8.8 (2012): 2949-2961.

Page 7: [1] Deleted DECLAN CLARKE 8/16/15 1:26 PM

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

Page 11: [2] Deleted DECLAN CLARKE 8/16/15 1:26 PM

This ubiquity suggests that, not only is allostery an essential component to understanding protein behavior in general, but it is also well suited to provide a conceptual framework for understanding many of the conserved regions found in proteins. Such conservation patterns are increasingly coming to light in the age of next-generation sequencing. As such, higher-throughput approaches for identifying potentially allosteric residues are needed to meet the high-throughput data generated through deep sequencing of both human and non-human genomes.

In order to

Page 11: [3] Deleted DECLAN CLARKE 8/16/15 1:26 PM

Notably, various methods for predicting allosteric hotspots have been described previously.

Page 11: [4] Moved to page 12 (Move #3)DECLAN CLARKE 8/16/15 1:26 PM

MD and NMR are some of the most common means of studying allostery and dynamic behavior. However, these methods have limitations when studying large and diverse protein datasets. MD is computationally expensive and impractical when studying large numbers of proteins. NMR structure determination is extremely labor-intensive and better suited to certain classes of structures or dynamics. In addition, NMR structures constitute a relatively small fraction of structures currently available.

**Page 11: [5] Deleted** **DECLAN CLARKE** **8/16/15 1:26 PM**

Given the limitations in applying MD, NMR, or related methods to large numbers of proteins, there remains a need to evaluate dynamic behavior in a systemized way across many proteins, while simultaneously capturing residues on both the surface and the interior.

**Page 11: [6] Deleted** **DECLAN CLARKE** **8/16/15 1:26 PM**

a large number of the alternative conformations available in the PDB. There is now a great deal of redundancy in folds and proteins: there are

**Page 11: [7] Deleted** **DECLAN CLARKE** **8/16/15 1:26 PM**

for which alternative crystal structures are available. This redundancy opens the door to large-scale analyses aimed at

**Page 11: [8] Deleted** **DECLAN CLARKE** **8/16/15 1:26 PM**

heterogeneity and potential allosteric behavior on a database-level scale. As such, we integrate data from the large number of X-ray crystal structures in the PDB to identify instances of these distinct conformations, which are then used as the raw material for identifying residues that may be important in the context of allosteric behavior.
We then

**Page 11: [9] Deleted** **DECLAN CLARKE** **8/16/15 1:26 PM**

. The critical residues identified (especially those which are interior and, to a lesser extent, on the surface) are shown

**Page 11: [10] Deleted** **DECLAN CLARKE** **8/16/15 1:26 PM**

HGMD was used in order to identify known disease-causing variants. We