Title:

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

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 12/8/15 3:18 PM
Deleted:

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: Identifying allosteric hotspots with dynamics: application to conservation in deep sequencing

DECLAN CLARKE 12/8/15 3:18 PM
Formatted: Font:Not Bold

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 unavailable. Allostery may often provide the missing explanatory link. We use models of protein conformational change to identify allosteric residues by predicting essential surface cavities or information flow bottlenecks, and we develop a software tool (stress.molmovdb.org) that enables users to perform this analysis on their own proteins of interest. Though fundamentally 3D-structural in nature, this software is computationally fast, thereby allowing us to run it across the PDB and to evaluate general properties of the predicted allosteric residues, which tend to be conserved over long and short evolutionary time scales. We highlight examples in which allosteric residues can help explain poorly understood disease-associated variants.

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: protein

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: protein

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: strong

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: , in many cases, we cannot find

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: (such as the need to engage in protein-protein interactions or to achieve a close-packed hydrophobic core). Allosteric hotspots

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: Here, we

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: such

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: . In particular, we predict allosteric residues that can act as

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: While our tool is

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: it

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: still

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: and tractable. This allows

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: entire Protein Data Bank

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: large-scale

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: . We find that these

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: significantly

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: both

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: Finally, we

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: specific

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: previously

INTRODUCTION

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

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

Given the importance of allosteric regulation, as well as the role of allostery in imparting efficient functionality, several methods have been devised for the identification of likely allosteric residues. Conservation itself has been used, either in the context of conserved residues (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). 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 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).

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).

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). Molecular dynamics (MD) and network analyses have been used to identify interior residues that may function as allosteric bottlenecks (Csermely et al., 2013; Gasper et al., 2012; Rousseau and Schymkowitz, 2005; Sethi et al., 2009; Vanwart et al., 2012). Along similar lines, Ghosh et al. have taken a novel approach of combining MD and network principles to characterize allosterically important inter-domain communication in methionyl tRNA synthetase (Ghosh et al., 2008). 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 feasibly be investigated), computational demands, or the class of residues to which the method is tailored (surface or interior). Using models of protein conformational change, we identify both surface and interior residues that may act as essential allosteric regions in a computationally tractable manner, thereby enabling high-throughput analysis. This framework directly incorporates information regarding protein structure and dynamics, and it is applied to proteins throughout the PDB (Berman et al., 2000) that exhibit conformational change. The relatively greater conservation of the residues identified (both across species and amongst humans) may help to elucidate many of the otherwise poorly understood regions in proteins. In a similar vein, several of our identified sites correspond to human disease loci for which no clear mechanism for pathogenesis had previously been proposed. Finally, our framework (termed STRESS, for STRucturally-identified ESSential residues) is made available through a tool to enable users to submit their own structures for analysis.

RESULTS

Identifying Potential Allosteric Residues

Allosteric residues at the surface generally play a regulatory role that is fundamentally distinct from that of allosteric residues within the protein interior. While surface residues may often constitute the sources or sinks of allosteric signals, interior residues act to transmit such signals. We use models of protein conformational change in an attempt to identify both classes of residues (Figure 1). Throughout, we term these potential allosteric residues at the surface and interior "surface-critical" and "interior-critical" residues, respectively. Critical residues are first identified in a set of 12 well-studied canonical systems (see Figure S1, as well as Table S1 for rationale), and they are then identified on a large scale across hundreds of distinct proteins.

Identifying Surface-Critical Residues

Allosteric ligands often act by binding to surface cavities and modulating protein conformational dynamics. The surface-critical residues, some of which may act as latent ligand binding sites and active sites, are first identified by finding cavities using Monte Carlo simulations to probe the surface with a flexible ligand (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 interfere 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

DECLAN CLARKE 12/8/15 3:18 PM
Formatted: Indent: First line: 0"

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: . ... [1]

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: for which both the *holo* and *apo* states are available (Table S1 and

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: .

unaffected (Figure 1A, bottom-right) earn lower scores. Further details are provided in SI Methods section 3.1-a.

This approach is a modified version of the binding leverage framework introduced by Mitternacht and Berezovsky (Mitternacht and Berezovsky, 2011). The main modifications 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). We find that this modified approach results in an average of ~2 distinct sites per domain (Figure 2A). The distribution for distinct sites within entire complexes 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 large-scale motions.

Dynamical Network Analysis to Identify Interior-Critical Residues

The binding leverage framework described above is intended to capture hotspot regions at the protein surface, but the Monte Carlo search employed is *a priori* excluded from the protein interior. Allosteric residues often act within the protein interior by

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: (see SI Methods).

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: (see SI Methods).

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: S2

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: ; see SI Methods for details on defining distinct sites).

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: 60

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: Tables S2 and S3

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: supplementary note "Capturing Known Ligand-Binding Sites").

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: S4

DECLAN CLARKE 12/8/15 3:18 PM
Formatted: Font:Arial, 11 pt, Bold

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: . ... [2]

functioning as essential 'bottlenecks' within the communication pathways between distal regions. An allosteric signal transmitted from one region to another may conceivably take various alternative routes, but many of these routes can share a common set of residues. The removal of such a common set of residues can result in the loss of many or all of the available routes for allosteric signal transmission, thereby making these residues essential information flow bottlenecks.

To identify bottlenecks, the protein is first modeled as a network, wherein residues represent nodes and edges represent contacts between residues (in much the same way that the protein is modeled as a network in constructing anisotropic network models, see below). In this regard, the problem of identifying interior-critical residues is reduced to a problem of identifying nodes that participate in network bottlenecks (see Figure 1B and SI Methods section 3.1-b for details). Briefly, the network edges are first weighted by the correlated motions of contacting residues: a strong correlation in the motion between contacting residues implies that knowing how one residue moves better enables one to predict the motion of the other, thereby suggesting a strong information flow between the two residues. The weights are used to assign 'effective distances' between connecting nodes, with strong correlations resulting in shorter effective node-node distances.

Using the motion-weighted network, "communities" of nodes are identified using the Girvan-Newman formalism (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)

The implementations for finding both surface- and interior-critical residues have been made available to the scientific community through a new software tool, STRESS, which may be accessed at stress.molmovdb.org (Figure 3A). Users may specify a PDB to be analyzed, and the output provided constitutes the set of identified critical residues.

Obviating the need for long wait times, the algorithmic implementation of our software is highly efficient (Figures 3B and 3C). A typical protein of ~500 residues takes only about 30 minutes on a 2.6GHz CPU. Running times are also minimized by using a scalable server architecture that runs on the Amazon cloud (Figure 3D). A light front-end server handles incoming user requests, and more powerful back-end servers, which perform the calculations, are automatically and dynamically scalable, thereby ensuring that they can handle varying levels of demand both efficiently and economically.

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: S3 and S4

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: S5

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: Figure S6

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: S7

High-Throughput Identification of Alternative Conformations

Pronounced conformational change is an essential assumption within our framework for identifying potential allosteric residues. We use a generalized approach to systematically identify instances of alternative conformations within the PDB. We first perform multiple structure alignments (MSAs) across sequence-identical proteins that are pre-filtered to ensure structural quality. We then use the resultant pairwise RMSD values to infer distinct conformational states (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 number of dynamic proteins culled throughout the PDB, coupled with the high algorithmic efficiency of our critical residue search implementation, provide a

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: Figures S8 and S9

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: for details

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: S2C

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: S2D

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: 2E. Further summary statistics are provided in Figure S10

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: (Figure S11).

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: S12

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: . ... [3]

DECLAN CLARKE 12/8/15 3:18 PM
Formatted: Indent: First line: 0.5"

DECLAN CLARKE 12/8/15 3:18 PM
Formatted: Font:Times New Roman, No underline

means of evaluating general properties of these residues on a large scale. In particular, we measure their conservation, as evaluated both over long (inter-species) and short (intra-human) evolutionary timescales. Using a variety of conservation metrics and sources of data, we find that both surface-critical (Figures 4A-D) and interior-critical (Figures 4E-H) are consistently more conserved than non-critical residues. We emphasize that the signatures of conservation identified not only provide a means of rationalizing many of the otherwise poorly understood regions of proteins, but they also reinforce the functional importance of the residues believed to be allosteric.

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: 3A

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: 3E

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). 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 ($p < 2.2e-16$; negative conservation scores designate stronger conservation). Interior-critical residues exhibit a similar trend: the mean conservation score for interior-critical residues and non-critical residues with the same degree of burial is -0.179 and -0.102, respectively ($p=3.67e-11$).

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: 3B

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: 3F

Measures of Conservation Amongst Humans from Next-Generation Sequencing

We may also use sequenced human genomes and exomes to investigate conservation, as many constraints may be human-specific and active in more recent evolutionary history. In this context, commonly used metrics for evaluating conservation include minor allele frequency (MAF) and derived allele frequency (DAF). Low MAF or

DAF values are interpreted as signatures of deleteriousness, as purifying selection is prone to reduce the frequencies of harmful variants (see SI Methods section 3.3-b).

We find that 1000 Genomes (McVean et al., 2012) single-nucleotide variants (SNVs) that hit surface-critical residues tend to occur at lower DAF values (Figure 4C). Though not 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) in which 1000 Genomes SNVs hit these critical sites. Furthermore, the long tail extending to lower DAF values for surface-critical residues may suggest that only a subset of the residues in our prioritized binding sites is essential. However 1000 Genomes SNVs tend to hit interior-critical residues with significantly lower DAF values than non-critical residues (Figure 4G; see also Figure S4B).

Given the relatively small number of proteins to be hit by 1000 Genomes SNVs, we also analyzed data provided by the Exome Aggregation Consortium (ExAC, Cambridge MA 2015). ExAC provides sequence data for many more individuals, and the ExAC sequencing itself is performed at much higher coverage. Thus, using MAF as a conservation metric, we performed a similar analysis using this data. MAF distributions for surface- and non-critical residues in the same set of proteins are given in 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

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: 3C

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: 13A

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: (see SI Methods).

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: 3G

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: S13B

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: 3D

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 overall allele frequency distributions, one may also evaluate the *fraction* of rare alleles as a metric for measuring selective pressure. This fraction is defined as the ratio of the number of low-DAF or low-MAF (i.e., rare) non-synonymous SNVs to all non-synonymous SNVs in a given protein annotation (such as all surface-critical residues of the protein, for example; see SI Methods 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 in critical residues (See Figure 5).

Comparisons Between Different Models of Protein Motions

Conformational changes may be modeled using vectors connecting pairs of corresponding residues in crystal structures from alternative conformations (we term this approach "ACT", for "absolute conformational transitions"; 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

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: S14A

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: 3H

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: S14B

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: Figures S15 and S16 for 1000 Genomes and ExAC data, respectively

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: ").

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: (further details in SI Methods).

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, and we thus use ANMs as our primary means of inferring motions.

Using vectors from either ACTs or ANMs give the same general results in terms of the disparities in conservation between critical and non-critical residues. This method is thus general with respect to how motion vectors are defined (see 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 the concept of SNV deleteriousness: changes in amino acid composition at specific loci may be more or less likely to result in disease. SIFT (Ng and Henikoff, 2001) and PolyPhen (Adzhubei et al., 2010) are two tools for predicting such effects, and we evaluated these predictions for critical and non-critical residues hit by SNVs in ExAC. SNVs hitting critical residues exhibit significantly higher PolyPhen scores relative to non-critical residues, suggesting the potentially higher disease susceptibility at critical residues (Figure S6), though such 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

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: S17

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: Supplemental note "Modeling Protein Motions by Directly Using Displacement Vectors from Alternative Conformations"

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: S18

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: S19

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: 4A

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: 4

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: 4B

that coincide with critical residues. In addition, we identify Y328 as a surface-critical residue, which coincides with a disease-associated SNV from HGDM, despite the lack of confident predictions of deleteriousness by several widely used tools for predicting disease-associated SNVs, including PolyPhen (Adzhubei et al., 2010), SIFT (Ng and Henikoff, 2001), and SNPs&GO (Calabrese et al., 2009). Together, these results suggest that the incorporation of surface- and interior-critical residues introduces a valuable layer of annotation to the protein sequence, and may help to explain otherwise poorly understood disease-associated SNVs.

DISCUSSION & CONCLUSIONS

The same principles of energy landscape theory that dictate protein folding are integral to how proteins explore different conformations once they adopt their folded states. These landscapes are shaped not only by the protein sequence itself, but also by extrinsic conditions. Such external factors often regulate protein activity by introducing allosteric-induced changes, which ultimately reflect changes in the shapes and population distributions of the energetic landscape. In this regard, allostery provides an ideal platform from which to study protein behavior in the context of their energetic landscapes. To investigate allosteric regulation, and to simultaneously add an extra layer of annotation in the context of conservation patterns, an integrated framework to identify potential allosteric residues is essential. We introduce a framework to select such residues, using knowledge of conformational change.

When applied to many proteins with distinct conformational changes in the PDB, we investigate the conservation of potential allosteric residues in both inter-species and intra-human genomes contexts, and find that these residues tend to exhibit greater conservation in both cases. In addition, we identify several disease-associated variants for which plausible mechanisms had previously been unavailable, but for which allosteric mechanisms provide a plausible rationale.

Unlike the characterization of many other structural features, such as secondary structure assignment, residue burial, protein-protein interaction interfaces, disorder, and even stability, allostery inherently manifests in the context of dynamic behavior. It is only by considering protein motions and changes in these motions can a fuller understanding of allosteric regulation be realized. As such, MD and NMR are some of the most common means of studying allostery and dynamic behavior (Kornev and Taylor, 2015). However, these methods have limitations when studying large and diverse protein datasets. MD is computationally expensive and impractical when studying large numbers of proteins. NMR structure determination is extremely labor-intensive and better suited to certain classes of structures or dynamics. In addition, NMR structures constitute a relatively small fraction of structures currently available.

Despite these limitations in MD and NMR, allosteric mechanisms and signaling pathways may be conserved across many different but related proteins, suggesting that such computationally- or labor-intensive approaches for all proteins may not be entirely essential. Flock et al. have carefully demonstrated that the allosteric mechanisms responsible for regulating G proteins through GPCRs tend to be conserved (Flock et al., 2015). If allosteric mechanisms are similarly shared within other protein families, a

463 detailed analysis with methods such as MD or NMR on one member of a family may help
464 to elucidate the allosteric behavior for other members. Nevertheless, the degree to which
465 these mechanisms are indeed conserved within other groups of proteins is currently
466 unclear, so homology-based predictions of allosteric mechanisms are still not readily
467 available.

468 Investigations into representative families have also been enlightening in other
469 contexts. In one of the early studies employing network analysis, del Sol et al. conduct a
470 detailed study of several allosteric protein families (including GPCRs) to demonstrate
471 that residues important for maintaining the integrity of short paths within residue contact
472 networks are essential to enabling signal transmission between distant sites (del Sol et al,
473 2006). Notably, many of the key sites identified correspond to residues that had been
474 experimentally determined to be important for allostery. Another notable result in the
475 same work is that these key residues may become redistributed when the protein
476 undergoes conformational change, thereby changing optimal communication routes in
477 different conformations as a means of conferring different regulatory properties.

478 There are several notable implications of our database-scale analysis. Relative to
479 sequence data, allostery and dynamic behavior are far more difficult to evaluate on a
480 large scale. The framework described here enables one to evaluate dynamic behavior in a
481 systemized and efficient way across many proteins, while simultaneously capturing
482 residues on both the surface and within the interior. That this pipeline can be applied in a
483 high-throughput manner enables the investigation of system-wide phenomena, such as
484 the roles of potential allosteric hotspots in protein-protein interaction networks.
485 Knowledge of such sites across many proteins may also be used to identify the best

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: Thus, there

proteins and protein regions for which drugs should be engineered, as well as instances in which specific sequence variants are likely to have the greatest impact.

We emphasize that it is only by applying this framework over a database of many proteins can one search for significant disparities in conservation between sites believed to be important in allostery and the rest of the protein. Such general trends may not be apparent when studying a small number or specific classes of proteins. To our knowledge, this is the first study in which the conservation of potential allosteric sites has been measured across a large database of proteins.

The ability to leverage our framework in a high-throughput manner also better enables one to match structural features with the high-throughput data generated through deep sequencing. Full human genomes and exomes are being sequenced at an increasing pace, thereby providing an unprecedented window into conservation patterns that can be human-specific or active over short evolutionary timescales. These patterns increasingly serve as detailed signatures of selective constraints which may not only be missing in cross-species comparisons, but are also sometimes difficult to rationalize using static representations of protein structures alone.

We anticipate that, within the next decade, deep sequencing will enable structural biologists to study evolutionary conservation using sequenced human exomes just as routinely as cross-species alignments. Furthermore, intra-species metrics for conservation provide added value in that the confounding factors of cross-species comparisons are removed: different organisms evolve in different cellular and evolutionary contexts, and it can be difficult to decouple these different effects from one another. Cross-species metrics of protein conservation entail comparisons between proteins that may be very

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

METHODS

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

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: S9

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: methods

DECLAN CLARKE 12/8/15 3:18 PM
Deleted:

ACKNOWLEDGMENTS

DC acknowledges the support of the NIH Predoctoral Program in Biophysics (T32 GM008283-24). We thank Simon Mitternacht for sharing the original source code for binding leverage calculations, as well as Koon-Kiu Yan for helpful discussions and feedback. The authors would like to thank the Exome Aggregation Consortium and the

groups that provided exome variant data for comparison. A full list of contributing groups can be found at http://exac.broadinstitute.org/about

REFERENCES

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

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 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

DECLAN CLARKE 12/8/15 3:18 PM
Deleted:

Equilibrated Structures of Methionyl tRNA Synthetase Complexes. Biochemistry. *47*, 11398-11407.

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

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

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., Lu, S., Huang, Z., Liu, X., Mou, L., Luo, Y., Zhao, Y., Liu, Y., Chen, Z., Hou, T., et al. (2013). Allosite: A method for predicting allosteric sites. Bioinformatics *29*, 2357–2359.

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

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

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

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

Proteins. Science *322*, 438-442.

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

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

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

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

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

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

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

Panjkovich, A. and Daura, X. (2010). Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery. BMC Struct. Biol. *10*, 9.

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

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

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

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

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

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

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

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

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

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

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

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

CAPTIONS

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

DECLAN CLARKE 12/8/15 3:18 PM
Formatted: Hyperlink, No underline, Font color: Auto

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: FIGURE

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: affect

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. Performing local searching supported with hashing and implementing additional algorithmic optimizations for computational efficiency reduce running times considerably (in green), relative to a more naïve approach without optimization (in red). (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

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: on database-wide analyses

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: The distributions of the number conformations (i.e., "K") for domains and chains are given in (*C*) and (*D*), respectively. Only proteins for which K exceeds 1 (for chains) are included in our dataset of multiple conformations. (*E*) Distinct proteins in our dataset within the context of high-quality X-ray structures in the PDB that we structurally aligned. A set of distinct proteins is such that no pair shares more than 90% sequence identity

DECLAN CLARKE 12/8/15 3:18 PM
Formatted: Font:Not Bold

DECLAN CLARKE 12/8/15 3:18 PM
Formatted: Tabs: 1.44", Left

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: Figure 3.

capable of handling wide fluctuations in user demand. Amazon's Simple Queue Service is used to coordinate between user requests at the front end and the back-end compute nodes: when the front-end server receives a request, it adds the job to the queue, and back-end servers pull that job from the queue when ready. Source code is available through Github (github.com/gersteinlab/STRESS).

Figure 4. Multiple metrics and datasets reveal that critical residues tend to be conserved. Surface- and interior-critical residues (red) in phosphofructokinase (PDB 3PFK) are given in (*A*) and (*E*), respectively. Distributions of cross-species conservation scores, 1000 Genomes SNV DAF averages, and ExAC SNV MAF averages for surface- and non-critical residue sets are given in (*B*), (*C*), and (*D*), respectively. The same distributions corresponding to interior- and non-critical residue sets are given in (F), (G), and (H), respectively. In (C), means for surface- and non-critical sets are 9.10e-4 and 8.34e-4, respectively (p=0.309); corresponding means in (*D*) are 4.09e-04 and 2.26e-04, respectively (p=1.49e-3). In (*G*), means for interior- and non-critical sets are 2.82e-4 and 3.12e-3, respectively (p=1.80e-05); corresponding means in (*H*) are 3.08e-05 and 3.27e-04, respectively (p=7.98e-09). Statistics for panels (*B*) and (*F*) are given in the main text. N = 421, 32, 84, 517, 31, and 90 structures for panels B, C, D, F, G, and H, respectively. P-values are based on Wilcoxon-rank sum tests. See SI Methods for further details.

Figure 5: Critical residues are shown to be more conserved, as measured by the fraction of rare alleles. Protein regions with high fractions of *rare* variants are believed to be more sensitive to sequence variants than other regions, thereby explaining why such

DECLAN CLARKE 12/8/15 3:18 PM
Formatted: Tabs: 1.44", Left

DECLAN CLARKE 12/8/15 3:18 PM
Deleted: 4

variants occur infrequently in the population. Panels *(A)* and *(C)* show distributions for rare (low DAF) non-synonymous SNVs (taken from the 1000 Genomes dataset) in which the critical residues are defined to be the surface-critical *(A)* and interior-critical *(C)* residues. Panels *(B)* and *(D)* show distributions for rare (low MAF) non-synonymous SNVs (taken from the ExAC dataset) in which the critical residues are defined to be the surface-critical *(B)* and interior-critical *(D)* residues. For varying thresholds to define rarity, there are more structures in which the fraction of rare variants is higher in critical residues than in non-critical residues. Cases in which the fraction is equal in both categories are not shown. We consider all structures such that at least one critical and at least one non-critical residue 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.

Figure 7. Potential allosteric residues add a layer of annotation to structures in the context of disease-associated SNVs. The structure shown (*A*) is that of the fibroblast

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

DECLAN CLARKE 12/8/15 3:18 PM
Formatted: Font:Not Bold

DECLAN CLARKE 12/8/15 3:18 PM
Formatted: Tabs: 1.44", Left

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

For each *apo* structure within the canonical set of proteins, statistics relating surface-critical sites to known ligand-binding sites are reported. The surface of a given structure is defined to be the set of all residues that have a relative solvent accessibility of at least 50%, where relative solvent accessibility is evaluated using all heavy atoms in both the main-chain and side-chain of a given residue. Mean values are given in the bottom row. NACCESS is used to calculate relative solvent accessibility (Hubbard and Thornton, 1993) . *Column 1*: PDB IDs for each structure; *Column 2*: among these surface residues, the fraction that constitute surface-critical residues; *Column 3*: among surface residues, the fraction that constitute known ligand-binding residues (known ligand-binding residues are taken to be those within 4.5 Angstroms of the ligand in the *holo* structure; Table S1); *Column 4*: the Jaccard similarity between the sets of residues represented in

columns 2 and 3 (i.e., surface-critical and known-ligand binding residues), where values given in parentheses represent the expected Jaccard similarity, given a null model in which surface-critical and ligand-binding residues are randomly distributed throughout the surface (for each structure, 10,000 simulations are performed to produce random distributions, and the expected values reported here constitute the mean Jaccard similarity among the 10,000 simulations for each structure); *Column 5*: the number of distinct surface-critical sites identified in each structure; *Column 6*: the number of known ligand-binding sites in each structure; *Column 7*: the number of known ligand-binding sites which are positively identified within the set of surface-critical sites, where a positive match occurs if a majority of the residues in a surface-critical site coincide with the known ligand-binding site; *Column 8*: The fraction of ligand-binding sites captured is simply the ratio of the values in column 7 to those in column 6.

Page 6: [1] Deleted	**DECLAN CLARKE**	**12/8/15 3:18 PM**

Page 7: [2] Deleted	**DECLAN CLARKE**	**12/8/15 3:18 PM**

Page 10: [3] Deleted	**DECLAN CLARKE**	**12/8/15 3:18 PM**