ABSTRACT

The rapidly growing volumes of data produced by next-generation sequencing initiatives enable more in-depth analysis of protein conservation than previously possible. Deep sequencing across human populations may uncover protein regions under strong selective constraints, despite the absence of readily obvious structural mechanisms responsible for such constraints. Given the complex nature and general ubiquity of allosteric regulation (Gunasekaran et al, 2004; Tsai et al, 2014), allostery provides a rational framework to understanding many of these otherwise cryptic sites. We describe a workflow to identify structures that constitute alternative conformations within the PDB, and then use models of conformational change to select key residues that may mediate allostery or otherwise function as significant regulatory sites. We identify essential patches on surfaces, as well as essential residues which can function as information flow bottlenecks deeper inside proteins. Though many previously-described approaches for identifying allosteric residues entail computationally expensive methods (such as MD) or otherwise rely on less direct measures (such as conservation), our framework is computationally tractable and mechanistic in nature, in the sense that conformational change is directly included in the search for essential sites. Downstream analyses of these residues reveal that they tend to be conserved both across species and human populations. In addition, we introduce a web server to enable these calculations on user-submitted structures.

DECLAN CLARKE 7/28/15 12:09 AM
Formatted: Font:24 pt

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the …apidly growing volumes o ... [1]

INTRODUCTION

DECLAN CLARKE 7/28/15 12:09 AM
Formatted: Font:24 pt

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 "cryptic sites" may often be provided by considering the protein's dynamic behavior and multiple conformations.

In addition to the conformations exhibited by an ensemble of structures, 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 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, thereby rewiring essential pathways of information flow within the interior. Internal residues essential to the integrity of these communication networks constitute bottlenecks.

Given the importance of allosteric regulation, 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 SVMs (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).

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

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: of a protein…is dynamic in ... [2]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: in general have been given greater attention. …n allosteric mechanism may ... [3]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: Interestingly, deep sequencing has unearthed a class of conserved residues for which no clear structural constraints seem to be responsible for such conservation (i.e., "cryptic sites"). …iven the importance ... [4]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: The importance of protein dynamics to allosteric function has been exploited for identifying allosteric sites in several other studies. Specifically, ... [5]

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 communities-based 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 communities-based network analysis to identify essential 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 application (with respect to the subclass of protein studied), scale (the numbers of proteins which may be feasibly investigated), or the class of residues to which the method is tailored (surface or interior). We describe a framework to identify instances of alternative conformations for a diverse set of proteins, and apply this to the PDB. We then determine both surface and interior residues that may potentially serve as essential allosteric regions in a computationally tractable manner, thereby enabling high-throughput analysis. Further advantages of our method include the fact that it directly incorporates information regarding protein dynamics (as oppose to using less direct measures such as conservation, which we evaluate only after our predictions). We note that the residues identified tend to be conserved both across species and amongst humans, and many of these regions correspond to the cryptic sites previously discussed. In a similar manner, several of our identified sites correspond to human disease loci for which no clear mechanism had previously been proposed. Finally, our pipeline (termed STRESS, for STRucturally-identified ESSential residues) is made available through a web server to which users may submit their own structures for analysis.

RESULTS

Identification of Potential 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

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: the

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: Several groups have applied molecular

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: analysis in order

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: molecular dynamics with the same

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: elements

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: i.e.,

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: identified (i.e.,

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: . ... [6]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: method

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: entire

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: sites within the protein

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: could

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: affect the thermodynamic st ... [7]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: . Specifically, allosteric regu ... [8]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: is mechanistic in nature

DECLAN CLARKE 7/28/15 12:09 AM
Deleted:), it may

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: applied to a wide variety of proteins

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: on a mass scale

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: it simultaneously captures b ... [9]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: the interior residues which ... [10]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: to the public

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: hosted by our group, from

DECLAN CLARKE 7/28/15 12:09 AM
Formatted ... [11]

DECLAN CLARKE 7/28/15 12:09 AM
Formatted ... [12]

binding sites in the former category, or affected distal sites in the latter), interior residues generally act to transmit allosteric signals across large distances.

Modified Binding Leverage to Identify Critical Residues on the Surface

The modified version of the binding leverage framework for predicting potential allosteric residues on the surface (Mitternacht and Berezovsky, 2011, also detailed in Methods) entails a series of Monte Carlo searches to probe the surface (with the protein being represented with all heavy atoms) with a simulated ligand, thereby generating a series of candidate sites. This approach ultimately results in an average of ~2 distinct binding sites per domain (Fig. 2a; see Methods for the details on defining distinct sites).

Surface residues important to allosteric behavior may either be the allosteric ligand binding sites themselves or affected regions distal to the binding site. In order to evaluate the extent to which this method identifies sites of the former category, we studied the ligand-binding sites and active sites of 12 well-studied classical systems for which the crystal structures of both the *holo* and *apo* states are available. We find that, out of the 12 canonical systems, 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 meet the thresholds needed for defining a site 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 exhibit overlap with binding sites (Supp. Table 4).

Dynamical Network Analysis to Identify Critical Residues within the Interior

Binding leverage captures hotspot regions close to or at the protein surface, but the Monte Carlo search employed is *a priori* excluded from the protein interior. Thus, we apply communities-based network analyses to the protein complexes of our dataset to identify important internal residues. Such residues often act as communication pathways between distal surface sites. Modeling the protein structure as a network of interacting residues, an edge between a given pair of residues designates a mutual proximity of 4.5

DECLAN CLARKE 7/28/15 12:09 AM
Moved (insertion) [1]

DECLAN CLARKE 7/28/15 12:09 AM
Formatted: Font:(Default) Times New Roman, No underline

DECLAN CLARKE 7/28/15 12:09 AM
Formatted: Font:11 pt, Italic

DECLAN CLARKE 7/28/15 12:09 AM
Moved (insertion) [2]

DECLAN CLARKE 7/28/15 12:09 AM
Moved (insertion) [3]

Angstroms. Edges are weighted on the basis of the correlated movements between contacting residues (see Methods).

Web Server (STRESS)

(Currently under development by Shantao and undergrad student Richard Chang).

Models of Protein Conformational Change

High-Throughput Identification of Structures in Distinct Energetic Wells

As a first step toward culling a high-confidence set of alternative conformations, we perform multiple structure alignments (MSAs) across sequence-identical proteins, with the structures having been filtered to ensure quality (see Methods and Fig. 1 for details). 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 as a means of quantifying similarity (Supp. Fig. 3), and

we use RMSD in our downstream analyses. The fully-processed output for identifying high-confidence alternative conformations is provided in Supp. File 1. Finally, Supp. Fig. 4 showcases the conformational transitions we observe in a diverse array of biological contexts, such conformational changes associated with ligand binding, protein-protein or protein-nucleic acid interactions, changes in oxidation or oligomerization state, etc.

Comparisons Between Different Models of Protein Motions

We evaluated the extent to which the results described here may be sensitive to different models of conformational change. ANMs are simple and straightforward to apply on a database scale, and are thus used as our primary model of choice. However, directly using the displacement vectors between all corresponding pairs of residues within the two crystal structures of the alternative conformations gives the same general results (see Supp. Fig. 15 and Supplemental discussion). Thus, our method is general with respect to how motions are defined.

DECLAN CLARKE 7/28/15 12:09 AM
Moved (insertion) [4]

DECLAN CLARKE 7/28/15 12:09 AM
Formatted: Font:11 pt, Italic

DECLAN CLARKE 7/28/15 12:09 AM
Formatted: Font:11 pt, Italic

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: To build the high-confidence dataset of conformational changes, we ... [13]

DECLAN CLARKE 7/28/15 12:09 AM
Formatted ... [14]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: domains as well as …rotein ... [15]

DECLAN CLARKE 7/28/15 12:09 AM
Moved down [5]: Fig.

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: 2). ... [16]

DECLAN CLARKE 7/28/15 12:09 AM
Moved (insertion) [5] ... [17]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: to reduce the uncertainty as ... [18]

DECLAN CLARKE 7/28/15 12:09 AM
Moved up [1]: . ... [19]

DECLAN CLARKE 7/28/15 12:09 AM
Formatted ... [20]

DECLAN CLARKE 7/28/15 12:09 AM
Formatted ... [21]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: We…use RMSD in our ... [22]

DECLAN CLARKE 7/28/15 12:09 AM
Moved up [2]: 2a; see Methods for ... [24]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: ... [25]

DECLAN CLARKE 7/28/15 12:09 AM
Moved down [6]: Fig.

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: 4). In order to be consistent ... [26]

DECLAN CLARKE 7/28/15 12:09 AM
Moved up [3]: Some of the sites id ... [27]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: However, two factors sugg ... [28]

DECLAN CLARKE 7/28/15 12:09 AM
Moved (insertion) [7] ... [29]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: Briefly, this method first er ... [23]

DECLAN CLARKE 7/28/15 12:09 AM
Formatted ... [30]

DECLAN CLARKE 7/28/15 12:09 AM
Moved (insertion) [8] ... [31]

Conservation Analyses on Critical Residues

Conservation Across Species

Our identified critical 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 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

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. 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 the 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 residuesFig. , the median for surface-critical residues is substantially lower than that for non-critical residues. The relative shifts of

DECLAN CLARKE 7/28/15 12:09 AM
Formatted: Font:(Default) Arial, 18 pt, Not Italic

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: *of Surface Sites*

DECLAN CLARKE 7/28/15 12:09 AM
Formatted: Font:(Default) Arial, 11 pt

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: Residues that lie in our prioritized sites…tend to be more conserved, on a ... [32]

DECLAN CLARKE 7/28/15 12:09 AM
Moved (insertion) [9]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: , Wilcoxon rank sum

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: *1000 Genomes- and ExAC-Derived*

DECLAN CLARKE 7/28/15 12:09 AM
Formatted: Font:(Default) Arial, 11 pt

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: *of Surface Sites*

DECLAN CLARKE 7/28/15 12:09 AM
Formatted: Font:(Default) Arial, 11 pt

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: *Modern-Day*

DECLAN CLARKE 7/28/15 12:09 AM
Formatted ... [33]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: at the level of 0.5% …Fig. ... [34]

DECLAN CLARKE 7/28/15 12:09 AM
Moved (insertion) [6]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: in our dataset that are…hit ... [35]

DECLAN CLARKE 7/28/15 12:09 AM
Moved down [10]: Fig.

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: 13a

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: an analogous…analysis usi ... [36]

DECLAN CLARKE 7/28/15 12:09 AM
Moved down [11]: Fig.

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: 10)….Although the mean r ... [37]

DECLAN CLARKE 7/28/15 12:09 AM
Moved (insertion) [12]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: 10, left), there is a skew toward lower minor allele frequencies for critical residues relative to non-critical residues. We also point out that…the median minor ... [38]

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.

As a separate test, SIFT and PolyPhen scores of critical and non-critical residues hit by variants from the ExAC dataset were evaluated. Though no significant disparity was observed in SIFT scores (Supp. Fig. 11), variants hitting critical residues exhibit significantly higher PolyPhen scores relative to non-critical residues (Supp. Fig. 12; note that higher PolyPhen scores denote more damaging variants). Together, these results suggest the more deleterious nature of changes to critical residues.

Critical Residues

in the Context of Human Disease Variants

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 mechanism is considered. Such disease loci constitute examples of "cryptic sites", and that our framework helps to shed light on such regions for which plausible alternative mechanisms of pathogenicity are not readily available.

Fibroblast growth factor receptor is a case-in-point (2F and Supp. Table 6), variants in which have been linked to diseases that manifest in craniofacial defects. Dotted lines highlight cryptic sites. The incorporation of critical-surface or critical-

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: 14a (…= 0.0475 and p= 8.7… [39]
DECLAN CLARKE 7/28/15 12:09 AM
Deleted: the distribution of derived … [40]
DECLAN CLARKE 7/28/15 12:09 AM
Formatted … [41]
DECLAN CLARKE 7/28/15 12:09 AM
Deleted: for 1000 Genomes, or allel… [42]
DECLAN CLARKE 7/28/15 12:09 AM
Moved (insertion) [10] … [43]
DECLAN CLARKE 7/28/15 12:09 AM
Deleted: for…different thresholds fo… [44]
DECLAN CLARKE 7/28/15 12:09 AM
Deleted: . Finally, we note that …IF… [46]
DECLAN CLARKE 7/28/15 12:09 AM
Formatted … [45]
DECLAN CLARKE 7/28/15 12:09 AM
Moved (insertion) [11] … [47]
DECLAN CLARKE 7/28/15 12:09 AM
Deleted: … [48]
DECLAN CLARKE 7/28/15 12:09 AM
Deleted: 8 and Supp. Table 7, respec… [53]
DECLAN CLARKE 7/28/15 12:09 AM
Moved up [7]: … [54]
DECLAN CLARKE 7/28/15 12:09 AM
Deleted: ***Conservation of Interior***
DECLAN CLARKE 7/28/15 12:09 AM
Moved up [9]: Figs.
DECLAN CLARKE 7/28/15 12:09 AM
Formatted … [49]
DECLAN CLARKE 7/28/15 12:09 AM
Deleted: ***Across Species*** … [50]
DECLAN CLARKE 7/28/15 12:09 AM
Moved down [13]: The distributio… [51]
DECLAN CLARKE 7/28/15 12:09 AM
Deleted: … [52]
DECLAN CLARKE 7/28/15 12:09 AM
Formatted … [55]
DECLAN CLARKE 7/28/15 12:09 AM
Deleted: . Given that our entire sch… [56]
DECLAN CLARKE 7/28/15 12:09 AM
Moved up [8]: ANMs are simple an… [57]
DECLAN CLARKE 7/28/15 12:09 AM
Deleted: … [58]
DECLAN CLARKE 7/28/15 12:09 AM
Formatted … [59]
DECLAN CLARKE 7/28/15 12:09 AM
Deleted: Within our dataset of high-… [60]
DECLAN CLARKE 7/28/15 12:09 AM
Deleted: A case-in-point is provided… [61]
DECLAN CLARKE 7/28/15 12:09 AM
Moved up [12]: Fig.
DECLAN CLARKE 7/28/15 12:09 AM
Deleted: Fig. 9…, variants in which… [62]

interior residues adds an additional layer of annotation to the protein sequence, and may thus help to explain otherwise poorly understood disease variants.

DISCUSSION & CONCLUSIONS

The same principles of energy landscape theory that dictate protein folding are essential to understanding how folded proteins explore different conformational states. As in the case for folding, 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 topology 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. 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 important role played by conformational change in the vast majority of well-studied proteins, many of which have been investigated as a result of their significance in disease. Some proteins captured in our pipeline of alternative conformations may not exhibit allosteric behavior as part of their native functionality within cells, but the multiple energetic minima captured in their crystal structures may nevertheless be exploited for protein engineering [[cite C.J. Wilson, others]] or in pharmaceutical contexts [[cite]].

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 not only labor-intensive and best suited to specific classes of structures or dynamics (such as those with greater

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: that fall in high binding-leverage sites or constitute high-betweenness loci in dynamic representations of the protein structure (i.e., our critical residues)

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: these critical sites

DECLAN CLARKE 7/28/15 12:09 AM
Moved up [4]:

Web Server (STRESS)
(Currently under development by Shantao and undergrad student Richard Chang).

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: Finally, as we have done for HGMD SNVs, we also searched the NCBI ClinVar database (Landrum et al, 2014) for instances in which our identified critical residues coincide with disease location ... [63]

DECLAN CLARKE 7/28/15 12:09 AM
Formatted: Font:11 pt, Italic

DECLAN CLARKE 7/28/15 12:09 AM
Formatted: Font:24 pt

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: have emerged as

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: in order to regulate needed ... [64]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: , such as post-translational ... [65]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: a sensible

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: Understanding allosteric si ... [66]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: In addition, we note that some

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: do

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: Molecular dynamics (

DECLAN CLARKE 7/28/15 12:09 AM
Deleted:)

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: Notably,

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: very

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: ,

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: is thus

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: Like MD, NMR yields imp ... [67]

disordered content, or motions that operate on different time scales), but in addition, they constitute a relatively small fraction of the available structures.

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 at once. This would also provide a means of better characterizing the large number of variants that have been shown to be deleterious through next-generation sequencing initiatives. Such a database-scale approach is also much easier to exploit in studies focused on large networks of protein-protein interactions.

A database-scale approach necessitates careful and appropriate processing of the available structural data. There is now a great deal of redundancy in folds and proteins: 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 potential allosteric behavior on a database-level scale.

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 introduce a hybrid method to identify essential residues at the surface and the protein interior that leverages knowledge of conformational heterogeneity. To identify potential allosteric residues closer to the protein surface, we describe a modified version of the binding leverage method developed by Mitternacht and Berezovsky. Heavy atoms within the protein are included when searching the surface for sites in which the introduction of a ligand could strongly perturb conformational changes, thereby finding sites that more closely reflect cavities in the protein topology. Secondly, after these sites identified, we use a formalism originally used in the context of protein folding (the energy gap [[cite]]), in order 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 search for residues that may act as bottlenecks between modules within the protein structure. As with the modified binding leverage approach, information regarding conformational change is used in this network-

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: (currently about 10%).

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: This type of investigation applied to many proteins simultaneously…also ... [68]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: data in …he available struc ... [69]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: Thus, motivated by the idea that large differences in shape correspond to distinct conformations that occupy different energetic wells (Fig. 2), we describe and implement a pipeline for the identification of structures in distinct conformations using a statistical formalism. In doing so, ... [70]

DECLAN CLARKE 7/28/15 12:09 AM
Moved down [14]: Users may submit protein structures to our server (STRESS) in order to perform their own analyses for identifying essential residues at the surface or within the interior. We emphasize that, as next-generation sequencing initiatives continue to provide a clearer picture of conservation at the residue level, structural biologists will increasingly find unexplained regions under strong selection.

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: Our server readily enables the user to probe their own protein for potential allosteric regions, thereby helping to shed light on many…of these distinctregions. ... [71]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: We employ a…dynamical ... [72]

based analysis: edges within the network of interacting residues (and modules) are weighted to reflect dynamic behavior.

Thus, we emphasize that, while many previous studies use sequence characteristics or biophysical properties of individual amino acids to investigate allostery, focus on only the interior or surface residues, or may otherwise be restricted to a small number of proteins, we work on many proteins simultaneously within a generalized framework to explicitly incorporate dynamic behavior in order to identify potentially allosteric residues on both the surface and within the interior.

We Our method is motivated to find many of the so-called cryptic elements in protein structures. investigate the conservation of our critical residues in both inter-species and intra-human genomes contexts. The critical residues identified (especially those which are interior and, to a lesser extent, on the surface) are shown to exhibit greater conservation in both contexts, suggesting that amino acid changes at these critical sites are be more deleterious than changes in other parts of the protein.

HGMD was used in order to identify known disease-causing variants, and we found that several disease SNVs hit our identified residues. We identify several disease variants for which plausible mechanisms had previously been unavailable, but for which allosteric mechanisms provide a plausible rationale.

Given that allostery has previously been studied in the context of individual proteins, there are several notable implications of our database-scale analysis. That this pipeline can be applied en masse suggests avenues for future studies, including applications to protein-protein interaction networks, or guiding experimental studies to prioritize residues that are candidates for allosteric behavior (cite Rama Ranganathan, others). Knowledge of predicted allosteric sites across many proteins may be used to identify the best proteins for which drugs should be engineered, as well as instances in which specific sequence variants are likely to have the greatest impact.

Users may submit protein structures to our server (STRESS) in order to perform their own analyses for identifying essential residues at the surface or within the interior. We emphasize that, as next-generation sequencing initiatives continue to provide a clearer picture of conservation at the residue level, structural biologists will increasingly

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: finds…residues (that are b... [73]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: use a *mechanistic* approach for identifying…both the surface and with... [74]

DECLAN CLARKE 7/28/15 12:09 AM
Moved (insertion) [15]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: Thus, we…investigate the ... [75]

DECLAN CLARKE 7/28/15 12:09 AM
Moved up [15]: Our method is motivated to find many of the so-called cryptic elements in protein structures.

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: any …nown disease-causin... [76]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: That we achieved compelling results suggests that the level of coarse graining (i.e., in X-ray crystal structures and using ANMs instead of MD) was low enough to still recapitulate biologically interesting findings. …hat this pipeline can be ap... [77]

DECLAN CLARKE 7/28/15 12:09 AM
Moved (insertion) [14]

find unexplained regions under strong selection. The framework described herein, as implemented in this server, readily enables the search for potential allosteric regions.

METHODS

DECLAN CLARKE 7/28/15 12:09 AM
Formatted: Font:24 pt

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.

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: -

DECLAN CLARKE 7/28/15 12:09 AM
Formatted: Font:(Default) Times New Roman, 12 pt, Not Bold

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.

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: structure

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: generated

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: The K-values for MSAs, as well as the motivating conceptual framework, are summarized in Fig 2. About 3000 different domains had a K-value of 1 (i.e., one conformation identified), whereas the K-values of close to 2000 domains exceed 1 (these exhibit multiple conformations, Fig. 2C). For proteins, close to 8000 had a K-value of 1, and about 1000 proteins had K-values that exceed 1. When performing K-means clustering with the gap statistic, very similar results were obtained when clustering structures on the basis of pairwise RMSD or pairwise Q_H (Supp. Fig. 3), so we use RMSD in our downstream analyses. ... [78]

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]]).

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: of the N MSAs

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. This method is motivated by the observation that 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-

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: is given here. Hundreds or thousands of

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 origi, 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 7/28/15 12:09 AM
Deleted: originally outlined by Mitternacht and Berezovsky

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.

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: potential

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: by Mitternacht and Berezovsky

DECLAN CLARKE 7/28/15 12:09 AM
Deleted:

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: (as well as individual proteins and standard PDBs) for several well-annotated allosteric and ligand-binding proteins [[list]]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: Analyzed more proteins as gold standard (from several refs). Results are provided on server.

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.

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: anisotropic network models.

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

DECLAN CLARKE 7/28/15 12:09 AM
Deleted:). A

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: between residues suggests

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: flow (see earlier discussion), and would thus result

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: are

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: thus

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: (thus

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: placing

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: in a short path which

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: thus more essential

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: intra-protein communicatio ... [79]

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: calculated. Residues that are

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: the

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: interactions

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: pairs of interacting

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: are assigned

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: be

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: class interior

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: . Edge betweenness is defined as the total sum of shortest paths

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: which that edge is involved, with path lengths equal to the sum of edge weights (see Sethi et al, 2009 for a more detailed discussion).

DECLAN CLARKE 7/28/15 12:09 AM
Moved (insertion) [13]

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).

FIGURE CAPTIONS

DECLAN CLARKE 7/28/15 12:09 AM
Deleted:

DECLAN CLARKE 7/28/15 12:09 AM
Formatted: Font:24 pt

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).

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: and in Fig X

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

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: . Known biological ligands are shown in white VDW rendering

DECLAN CLARKE 7/28/15 12:09 AM
Formatted: Font:Bold

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: Database-wide distributions

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: derived allele frequency (DAF) values of surface critical residues (red) and non-critical residues (blue); **c)** Corresponding distributions of

DECLAN CLARKE 7/28/15 12:09 AM
Formatted: Font:Bold

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: critical residues (red) and non-critical residues (blue) **d**

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: **e)** Database-wide distributions of DAF values of interior critical residues (red) and non-critical residues (blue) **f)** Corresponding distributions

DECLAN CLARKE 7/28/15 12:09 AM
Deleted:

DECLAN CLARKE 7/28/15 12:09 AM
Deleted: red) and non-critical residues (blue

correspond with HGMD loci (green), and HGMD SNVs in non-critical residues (orange) are shown in VDW spheres.

REFERENCES

DECLAN CLARKE 7/28/15 12:09 AM
Formatted: Font:24 pt

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.

DECLAN CLARKE 7/28/15 12:09 AM
Moved (insertion) [16]

DECLAN CLARKE 7/28/15 12:09 AM
Moved up [16]:
Huang, Zhimin, et al. "ASD: a comprehensive database of allosteric proteins and modulators." Nucleic acids research 39.suppl 1 (2011): D663-D669.

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.

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 1: [1] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted DECLAN CLARKE 7/28/15 12:09 AM

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted DECLAN CLARKE 7/28/15 12:09 AM

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted DECLAN CLARKE 7/28/15 12:09 AM

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted DECLAN CLARKE 7/28/15 12:09 AM

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted DECLAN CLARKE 7/28/15 12:09 AM

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted DECLAN CLARKE 7/28/15 12:09 AM

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted DECLAN CLARKE 7/28/15 12:09 AM

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted DECLAN CLARKE 7/28/15 12:09 AM

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted DECLAN CLARKE 7/28/15 12:09 AM

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted DECLAN CLARKE 7/28/15 12:09 AM

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted DECLAN CLARKE 7/28/15 12:09 AM

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted DECLAN CLARKE 7/28/15 12:09 AM

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted DECLAN CLARKE 7/28/15 12:09 AM

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 1: [1] Deleted DECLAN CLARKE 7/28/15 12:09 AM

As the pace of novel fold discovery has slowed, and as structure determination has become more routine, there is a growing degree of redundancy within the PDB, whereby a given fold or protein is represented by multiple structures. Coincident with these trends, the

Page 2: [2] Deleted DECLAN CLARKE 7/28/15 12:09 AM

of a protein

Page 2: [2] Deleted DECLAN CLARKE 7/28/15 12:09 AM

of a protein

Page 2: [2] Deleted DECLAN CLARKE 7/28/15 12:09 AM

of a protein

Page 2: [2] Deleted DECLAN CLARKE 7/28/15 12:09 AM

of a protein

Page 2: [2] Deleted DECLAN CLARKE 7/28/15 12:09 AM

of a protein

Page 2: [2] Deleted DECLAN CLARKE 7/28/15 12:09 AM

of a protein

Page 2: [2] Deleted DECLAN CLARKE 7/28/15 12:09 AM

of a protein

Page 2: [2] Deleted DECLAN CLARKE 7/28/15 12:09 AM

of a protein

Page 2: [2] Deleted DECLAN CLARKE 7/28/15 12:09 AM

of a protein

Page 2: [3] Deleted DECLAN CLARKE 7/28/15 12:09 AM

in general have been given greater attention.

Page 2: [3] Deleted DECLAN CLARKE 7/28/15 12:09 AM

in general have been given greater attention.

Page 2: [3] Deleted DECLAN CLARKE 7/28/15 12:09 AM

in general have been given greater attention.

Page 2: [3] Deleted DECLAN CLARKE 7/28/15 12:09 AM

in general have been given greater attention.

Page 2: [3] Deleted DECLAN CLARKE 7/28/15 12:09 AM

in general have been given greater attention.

Page 2: [3] Deleted DECLAN CLARKE 7/28/15 12:09 AM

in general have been given greater attention.

Page 2: [3] Deleted DECLAN CLARKE 7/28/15 12:09 AM

in general have been given greater attention.

Page 2: [3] Deleted DECLAN CLARKE 7/28/15 12:09 AM

in general have been given greater attention.

Page 2: [4] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Interestingly, deep sequencing has unearthed a class of conserved residues for which no clear structural constraints seem to be responsible for such conservation (i.e., “cryptic sites”).

Page 2: [4] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Interestingly, deep sequencing has unearthed a class of conserved residues for which no clear structural constraints seem to be responsible for such conservation (i.e., “cryptic sites”).

Page 2: [4] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Interestingly, deep sequencing has unearthed a class of conserved residues for which no clear structural constraints seem to be responsible for such conservation (i.e., “cryptic sites”).

Page 2: [4] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Interestingly, deep sequencing has unearthed a class of conserved residues for which no clear structural constraints seem to be responsible for such conservation (i.e., “cryptic sites”).

Page 2: [4] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Interestingly, deep sequencing has unearthed a class of conserved residues for which no clear structural constraints seem to be responsible for such conservation (i.e., “cryptic sites”).

Page 2: [4] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Interestingly, deep sequencing has unearthed a class of conserved residues for which no clear structural constraints seem to be responsible for such conservation (i.e., “cryptic sites”).

Page 2: [4] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Interestingly, deep sequencing has unearthed a class of conserved residues for which no clear structural constraints seem to be responsible for such conservation (i.e., “cryptic sites”).

Page 2: [4] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Interestingly, deep sequencing has unearthed a class of conserved residues for which no clear structural constraints seem to be responsible for such conservation (i.e., “cryptic sites”).

Page 2: [5] Deleted DECLAN CLARKE 7/28/15 12:09 AM

The importance of protein dynamics to allosteric function has been exploited for identifying allosteric sites in several other studies. Specifically, normal

Page 2: [5] Deleted DECLAN CLARKE 7/28/15 12:09 AM

The importance of protein dynamics to allosteric function has been exploited for identifying allosteric sites in several other studies. Specifically, normal

Page 2: [5] Deleted DECLAN CLARKE 7/28/15 12:09 AM

The importance of protein dynamics to allosteric function has been exploited for identifying allosteric sites in several other studies. Specifically, normal

Page 3: [6] Deleted DECLAN CLARKE 7/28/15 12:09 AM

With the objective of focusing on those proteins that may undergo conformational change as part of their allosteric behavior, we develop and apply a general

Page 3: [7] Deleted DECLAN CLARKE 7/28/15 12:09 AM

affect the thermodynamic stability of these conformational states

Page 3: [8] Deleted DECLAN CLARKE 7/28/15 12:09 AM

. Specifically, allosteric regulation is usually imparted by a small subset of residues within a given protein: these may constitute a set of residues on the protein surface (such as an effector binding site) or a channel of residues within the interior that are responsible for linking distal sites (i.e., bottleneck residues). Our objective is to identify both of these classes of residues within a unified study. Several of our identified sites correspond to human disease loci for which no clear mechanism had previously been proposed.

Several of the

Page 3: [9] Deleted DECLAN CLARKE 7/28/15 12:09 AM

it simultaneously captures both surface sites and

Page 3: [10] Deleted DECLAN CLARKE 7/28/15 12:09 AM

the interior residues which may bridge such sites.

Page 3: [11] Formatted DECLAN CLARKE 7/28/15 12:09 AM

Font:24 pt

Page 3: [12] Formatted DECLAN CLARKE 7/28/15 12:09 AM

Indent: First line: 0"

Page 5: [13] Deleted DECLAN CLARKE 7/28/15 12:09 AM

To build the high-confidence dataset of conformational changes, we use a generalized approach to leverage the wealth of data in the PDB for systematically identifying proteins that occupy alternative energetic wells.

Page 5: [14] Formatted DECLAN CLARKE 7/28/15 12:09 AM

Indent: First line: 0"

Page 5: [15] Deleted DECLAN CLARKE 7/28/15 12:09 AM

domains as well as

Page 5: [15] Deleted DECLAN CLARKE 7/28/15 12:09 AM

domains as well as

Page 5: [15] Deleted DECLAN CLARKE 7/28/15 12:09 AM

domains as well as

Page 5: [15] Deleted DECLAN CLARKE 7/28/15 12:09 AM

domains as well as

Page 5: [15] Deleted DECLAN CLARKE 7/28/15 12:09 AM

domains as well as

Page 5: [16] Deleted DECLAN CLARKE 7/28/15 12:09 AM

2).

After performing multiple structure alignments for each sequence-identical group of proteins, we use

Page 5: [16] Deleted DECLAN CLARKE 7/28/15 12:09 AM

2).

After performing multiple structure alignments for each sequence-identical group of proteins, we use

Page 5: [16] Deleted DECLAN CLARKE 7/28/15 12:09 AM

2).

After performing multiple structure alignments for each sequence-identical group of proteins, we use

Page 5: [17] Moved from page 5 (Move #5)DECLAN CLARKE 7/28/15 12:09 AM

Fig.

Page 5: [18] Deleted DECLAN CLARKE 7/28/15 12:09 AM

to reduce the uncertainty associated with limited crystallographic resolution

Page 5: [18] Deleted DECLAN CLARKE 7/28/15 12:09 AM

to reduce the uncertainty associated with limited crystallographic resolution

Page 5: [19] Moved to page 4 (Move #1)DECLAN CLARKE 7/28/15 12:09 AM

.

Modified Binding Leverage to Identify Critical Residues on the Surface

Page 5: [20] Formatted DECLAN CLARKE 7/28/15 12:09 AM

Font:(Default) Times New Roman, No underline

Page 5: [21] Formatted DECLAN CLARKE 7/28/15 12:09 AM

Font:11 pt, Italic

Page 5: [22] Deleted DECLAN CLARKE 7/28/15 12:09 AM

We

Page 5: [22] Deleted DECLAN CLARKE 7/28/15 12:09 AM

We

Page 5: [22] Deleted DECLAN CLARKE 7/28/15 12:09 AM

We

Page 5: [22] Deleted DECLAN CLARKE 7/28/15 12:09 AM

We

Page 5: [22] Deleted DECLAN CLARKE 7/28/15 12:09 AM

We

Page 5: [23] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Briefly, this method first entails using a series of Monte Carlo simulations to probe the protein surface (with the protein being represented with all heavy atoms) with a simulated ligand, thereby generating a series of candidate sites. Each candidate site is then scored on the basis of the degree to which the occlusion (with the simulated ligand) disrupts the large-scale motions of the protein (Fig. 1, bottom left; see Methods). These motions are taken from anisotropic network models, but the results do not change drastically if we directly use the alternative conformations as given by the crystal structures (see "Comparisons Between Different Models of Protein Motions" in Results). The main modifications to the formalism described by Mitternacht and Berezovsky include the inclusion of heavy atoms in the protein during the Monte Carlo search, in addition to an automated means of thresholding the list of ranked sites to give a more selective set of candidate sites. This approach results in finding an average of ~2 distinct binding sites per domain (Fig.

Page 5: [24] Moved to page 4 (Move #2)DECLAN CLARKE 7/28/15 12:09 AM

2a; see Methods for the details on defining distinct sites).

Page 5: [25] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Surface residues important to allosteric behavior may either be the binding sites for allosteric ligands or the allosteric 'sinks' in signal transmission (that is, the sites affected by binding to a ligand in a distal region). In order to evaluate the extent to which this method identifies sites of the former category, we studied the ligand-binding sites and

active sites in a set of 12 well-studied systems for which the crystal structures of both the *holo* and *apo* states are available (Supp. Table 1; note that these 12 systems differ from the proteins highlighted in our alternative conformations analysis, as that set of proteins highlights our ability to identify alternative conformations in many different biological contexts, and not just ligand binding – see "Identifying Distinct Conformations in an Ensemble of Structures" in Methods and also Supp.

Page 5: [26] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

4). In order to be consistent with the original binding leverage study, these 12 constitute those used by Mitternacht and Berezovsky for which the *apo* and *holo* states are clear (Supp. Table 1).

We find that, out of the 12 canonical systems studied, we positively identify an average of 60% of the sites known to be directly involved in ligand or substrate binding. It has previously been shown that it is especially difficult to identify the sites in aspartate transcarbamoylase (Mitternacht and Berezovsky, 2011); excluding aspartate transcarbamoylase from this analysis results in finding an average of 65% of known biological sites. We note that these statistics are achieved by covering an average of 15% of proteins' residues (Supp. Table 2). For most proteins, selecting 15% of the residues is conservative -- more than 15% of the proteins' residues are involved in ligand or substrate binding for most proteins (Supp. Table 3).

Page 5: [27] Moved to page 4 (Move #3)DECLAN CLARKE **7/28/15 12:09 AM**

Some of the sites identified do not meet the thresholds needed for defining a site of known biological significance.

Page 5: [28] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

However, two factors suggest that such sites may still be significant in an allosteric context. First, we emphasize that 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. Such latent allosteric pockets may be useful in the context of drug development and targeting. 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).

Page 5: [29] Moved from page 7 (Move #7)DECLAN CLARKE 7/28/15 12:09 AM

Comparisons Between Different Models of Protein Motions

Page 5: [30] Formatted DECLAN CLARKE 7/28/15 12:09 AM

Font:11 pt, Italic

Page 5: [31] Moved from page 7 (Move #8)DECLAN CLARKE 7/28/15 12:09 AM

ANMs are simple and straightforward to apply on a database scale, and are thus used as our primary model of choice.

Page 6: [32] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Residues that lie in our prioritized sites

Page 6: [32] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Residues that lie in our prioritized sites

Page 6: [32] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Residues that lie in our prioritized sites

Page 6: [32] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Residues that lie in our prioritized sites

Page 6: [33] Formatted DECLAN CLARKE 7/28/15 12:09 AM

Font:(Default) Arial, 11 pt

Page 6: [33] Formatted DECLAN CLARKE 7/28/15 12:09 AM

Font:(Default) Arial, 11 pt

Page 6: [34] Deleted DECLAN CLARKE 7/28/15 12:09 AM

at the level of 0.5%

Page 6: [34] Deleted DECLAN CLARKE 7/28/15 12:09 AM

at the level of 0.5%

Page 6: [34] Deleted DECLAN CLARKE 7/28/15 12:09 AM

at the level of 0.5%

Page 6: [34] Deleted DECLAN CLARKE 7/28/15 12:09 AM

at the level of 0.5%

Page 6: [35] Deleted DECLAN CLARKE 7/28/15 12:09 AM

in our dataset that are

Page 6: [35] Deleted DECLAN CLARKE 7/28/15 12:09 AM

in our dataset that are

Page 6: [35] Deleted DECLAN CLARKE 7/28/15 12:09 AM

in our dataset that are

Page 6: [36] Deleted DECLAN CLARKE 7/28/15 12:09 AM

an analogous

Page 6: [36] Deleted DECLAN CLARKE 7/28/15 12:09 AM

an analogous

Page 6: [36] Deleted DECLAN CLARKE 7/28/15 12:09 AM

an analogous

Page 6: [36] Deleted DECLAN CLARKE 7/28/15 12:09 AM

an analogous

Page 6: [37] Deleted DECLAN CLARKE 7/28/15 12:09 AM

10).

Page 6: [37] Deleted DECLAN CLARKE 7/28/15 12:09 AM

10).

Page 6: [38] Deleted DECLAN CLARKE 7/28/15 12:09 AM

10, left), there is a skew toward lower minor allele frequencies for critical residues relative to non-critical residues. We also point out that

Page 6: [38] Deleted DECLAN CLARKE 7/28/15 12:09 AM

10, left), there is a skew toward lower minor allele frequencies for critical residues relative to non-critical residues. We also point out that

Page 7: [39] Deleted DECLAN CLARKE 7/28/15 12:09 AM

14a (

Page 7: [39] Deleted DECLAN CLARKE 7/28/15 12:09 AM

14a (

Page 7: [40] Deleted DECLAN CLARKE 7/28/15 12:09 AM

the distribution of derived

Page 7: [40] Deleted DECLAN CLARKE 7/28/15 12:09 AM

the distribution of derived

Page 7: [40] Deleted DECLAN CLARKE 7/28/15 12:09 AM

the distribution of derived

Page 7: [41] Formatted DECLAN CLARKE 7/28/15 12:09 AM

Font:Italic

Page 7: [42] Deleted DECLAN CLARKE 7/28/15 12:09 AM

for 1000 Genomes, or alleles with low minor allele frequency in ExC) in critical residues and non-critical residues for those proteins for which at least 1 critical residue and 1 non-critical residue is hit by an SNV.

Page 7: [42] Deleted DECLAN CLARKE 7/28/15 12:09 AM

for 1000 Genomes, or alleles with low minor allele frequency in ExC) in critical residues and non-critical residues for those proteins for which at least 1 critical residue and 1 non-critical residue is hit by an SNV.

Page 7: [43] Moved from page 6 (Move #10)DECLAN CLARKE 7/28/15 12:09 AM

Fig.

Page 7: [44] Deleted DECLAN CLARKE 7/28/15 12:09 AM

for

Page 7: [44] Deleted DECLAN CLARKE 7/28/15 12:09 AM

for

Page 7: [44] Deleted DECLAN CLARKE 7/28/15 12:09 AM

for

Page 7: [45] Formatted DECLAN CLARKE 7/28/15 12:09 AM

Indent: First line: 0.5"

Page 7: [46] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Finally, we note that

Page 7: [46] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Finally, we note that

Page 7: [46] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Finally, we note that

Page 7: [46] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Finally, we note that

Page 7: [46] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Finally, we note that

Page 7: [46] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Finally, we note that

Page 7: [47] Moved from page 6 (Move #11)DECLAN CLARKE 7/28/15 12:09 AM

Fig.

Page 7: [48] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Identifying Internal Residues for Transmitting Allosteric Information Using Dynamical Network Analysis

The binding leverage framework described above captures hotspot regions close to or at the surface of the protein, but the Monte Carlo search employed is *a priori* excluded from the protein interior. Thus, motivated by previous studies focused on individual proteins, such as tRNA synthetase (Sethi et al, 2009), essential metabolic enzymes (Manley et al, 2013), and the HIV envelope glycoprotein (Sethi et al, 2013), we apply communities-based network analyses to the protein complexes of our dataset to identify important internal residues, in addition to residues that may be closer to the surface, and note that the identification of residues in the protein interior are often needed to serve as communication channels between surface sites.

The nodes of the network represent individual residues, and edges between these nodes are drawn between residues that lie within a mutual proximity of 4.5 Angstroms. The first step in this analysis is to define the communities within the network. A given "community" refers to a group of residues that are highly inter-connected, but which have minimal edges to residues outside the community (see Methods). We have tried applying both an information-theory based method (termed "Infomap", Rosvall et al, 2007) and the classical Girvan-Newman (GN) formalism (Girvan et al, 2002) to decompose networks of interacting protein residues into communities (see Methods), and find that GN is more appropriate (see Supplemental discussion). The results described here are thus based on the GN formalism.

In order to recapitulate the contributions of various edges to information flow over the course of conformational

Page 7: [48] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Identifying Internal Residues for Transmitting Allosteric Information Using Dynamical Network Analysis

The binding leverage framework described above captures hotspot regions close to or at the surface of the protein, but the Monte Carlo search employed is *a priori* excluded from the protein interior. Thus, motivated by previous studies focused on individual proteins, such as tRNA synthetase (Sethi et al, 2009), essential metabolic enzymes (Manley et al, 2013), and the HIV envelope glycoprotein (Sethi et al, 2013), we apply communities-based network analyses to the protein complexes of our dataset to identify important internal residues, in addition to residues that may be closer to the surface, and note that the identification of residues in the protein interior are often needed to serve as communication channels between surface sites.

The nodes of the network represent individual residues, and edges between these nodes are drawn between residues that lie within a mutual proximity of 4.5 Angstroms. The first step in this analysis is to define the communities within the network. A given "community" refers to a group of residues that are highly inter-connected, but which have minimal edges to residues outside the community (see Methods). We have tried applying both an information-theory based method (termed "Infomap", Rosvall et al, 2007) and the classical Girvan-Newman (GN) formalism (Girvan et al, 2002) to decompose networks of interacting protein residues into communities (see Methods), and find that GN is more appropriate (see Supplemental discussion). The results described here are thus based on the GN formalism.

In order to recapitulate the contributions of various edges to information flow over the course of conformational

Page 7: [49] Formatted DECLAN CLARKE 7/28/15 12:09 AM

Font:(Default) Arial, 11 pt

Page 7: [50] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Across Species

Residues which function as essential allosteric conduits of information in mediating signal transduction from one site of a protein to another are likely to be more conserved, on average, than residues which lie outside of such channels. Thus, as for the case with surface critical residues, we evaluate conservation of residues identified as critical by GN, and compare these conservation scores to those of non-critical residues with the same degree (Fig. 3F).

Again, each point in this plot represents the mean conservation score for all residues within one of two classes (the set of interior critical residues or the randomly-selected set of non-critical residues with the same degree) within a particular protein structure. The randomly-selected non-critical set of residues was chosen in a way such that, for each critical residue with degree K, a randomly-chosen non-critical residue with the same degree K was included in the set.

Page 7: [51] Moved to page 18 (Move #13)DECLAN CLARKE 7/28/15 12:09 AM

The distribution of non-critical residues shown is very much representative of the distribution observed when re-building the random set many times.

Page 7: [52] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Interior critical residues are generally found to be more conserved than non-critical residues with the same degree of burial: the average conservation score for interior critical residues is -0.179, whereas that for non-critical residues with the same degree is -0.102 ($p=3.67e-11$, Wilcoxon rank sum test).

1000 Genomes- and ExAC-Derived Conservation of Interior Residues Amongst Modern-Day Humans

The residues we identify as bottlenecks are shown to be under negative selection in the context of modern-day humans. Our analyses of variants identified from The 1000 Genomes shows that interior critical residues occur at sites with significantly lower DAF values (Fig. 3E). Similar results are obtained for ExAC: interior critical residues exhibit a significantly lower minor allele frequency than do non-critical residues (Supp. Fig. 10, right). We also performed an analysis of the potential shifts the distributions of the mean

DAF values using a two-sample Kolmogorov-Smirnov test for 1000 Genomes (p=8.9E-5, Supp. Fig. 13b) and ExAC variants (Supp. Fig. 14b, p=8.7E-5).

We examined the *fraction* of rare alleles in critical interior residues and non-critical residues for those proteins for which at least 1 critical residue and 1 non-critical residue is hit by an SNV. The corresponding results for 1000 Genomes and ExAC variants are given in Supp.

Page 7: [53] Deleted DECLAN CLARKE 7/28/15 12:09 AM

8 and Supp. Table 7, respectively. Using a rarity threshold of 0.5% (0.1%) with 1000 Genomes variants, the fraction of rare SNVs in critical residues exceeds that in non-critical residues for 10 (20) structures (in green), whereas the fraction of rare allele in non-critical surface residues exceeds that in critical residues for only 0 (1) structure (in gray). When measuring human-specific conservation using the *fraction* of rare alleles in ExAC, interior-critical residues are also more conserved than non-critical residues at varying thresholds for rarity (Supp. Table 7). 15.5% (41.1%) of the structures studied were such that the fraction of rare allele in critical residues exceeded that for non-critical residues using a threshold of 0.005 (0.001), whereas the opposite trend was observed in only 0% (3.3%) of structures.

SIFT and PolyPhen scores of interior critical residues hit by variants from the ExAC dataset exhibited trends similar to those observed for surface-critical residues: no significant difference was seen between interior critical and non-critical residues with respect to SIFT scores (Supp. Fig. 11, right). However, these critical residues were shown to exhibit significantly higher PolyPhen scores relative to non-critical residues (Supp. Fig. 12, right), suggesting that modifications to these critical residues are significantly more damaging.

Page 7: [54] Moved to page 5 (Move #7)DECLAN CLARKE 7/28/15 12:09 AM

Comparisons Between Different Models of Protein Motions

Page 7: [55] Formatted DECLAN CLARKE 7/28/15 12:09 AM

Font:11 pt, Italic

Page 7: [56] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Given that our entire scheme is based on an understanding of protein motions, we evaluated the extent to which the results may be sensitive to different models of conformational change.

Page 7: [57] Moved to page 5 (Move #8)DECLAN CLARKE 7/28/15 12:09 AM

ANMs are simple and straightforward to apply on a database scale, and are thus used as our primary model of choice.

Page 7: [58] Deleted DECLAN CLARKE 7/28/15 12:09 AM

As an alternative to ANMs, one may simply use the displacement vectors between all corresponding pairs of residues within the two crystal structures of the alternative conformations for a given protein. This more direct model of conformational change, which we term absolute conformational transitions (ACT), may be applied in a straightforward manner to single-chain proteins. When we use ACT to apply the modified binding leverage framework for such single-chain proteins, we observe that our surface critical residues are significantly more conserved than are non-critical residues (Supp. Fig. 15, left). The same trend is observed when ACT is applied in our dynamical network analysis for identifying interior critical residues (Supp. Fig. 15, right).

Thus, despite the different ways of modeling conformational change, we find that our conservation results hold for different models, thereby demonstrating that our method is general with respect to how motions are defined.

Critical Residues that Coincide with

Page 7: [59] Formatted DECLAN CLARKE 7/28/15 12:09 AM

Font:11 pt, Italic

Page 7: [60] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Within our dataset of high-confidence alternative conformations

Page 7: [60] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Within our dataset of high-confidence alternative conformations

Page 7: [60] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Within our dataset of high-confidence alternative conformations

Page 7: [60] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

Within our dataset of high-confidence alternative conformations

Page 7: [60] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

Within our dataset of high-confidence alternative conformations

Page 7: [60] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

Within our dataset of high-confidence alternative conformations

Page 7: [60] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

Within our dataset of high-confidence alternative conformations

Page 7: [60] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

Within our dataset of high-confidence alternative conformations

Page 7: [60] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

Within our dataset of high-confidence alternative conformations

Page 7: [60] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

Within our dataset of high-confidence alternative conformations

Page 7: [60] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

Within our dataset of high-confidence alternative conformations

Page 7: [61] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

A case-in-point is provided by a fibroblast

Page 7: [61] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

A case-in-point is provided by a fibroblast

Page 7: [62] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

Fig. 9

Page 7: [62] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

Fig. 9

Page 7: [62] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

Fig. 9

Page 7: [62] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

Fig. 9

Page 7: [62] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

Fig. 9

Page 8: [63] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Finally, as we have done for HGMD SNVs, we also searched the NCBI ClinVar database (Landrum et al, 2014) for instances in which our identified critical residues coincide with disease locations. The affected proteins generally match those identified in the HGMD analysis above, and results are given in Supplementary Files 6 and 7 (for surface and interior critical residues, respectively).

Page 8: [64] Deleted DECLAN CLARKE 7/28/15 12:09 AM

in order to regulate needed functions.

Page 8: [65] Deleted DECLAN CLARKE 7/28/15 12:09 AM

, such as post-translational modifications, interactions with ligands, substrates, or other proteins, or physiological conditions within the cell.

Page 8: [66] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Understanding allosteric signal transmission inevitably entails a consideration of the dynamic properties that generally accompany and are required for such allosteric behavior, as well as the identification of the residues that are essential for such behavior. Although not all allosteric proteins undergo conformational change (i.e., allosteric signals may sometimes be transmitted by changing the frequency of native-state fluctuations, rather than imparting large changes in conformational topology) [[cite Rodgers, Nussinov]] and not all conformational changes are associated with allostery [[cite ex: Calmodulin]], it is frequently the case that allosteric behavior is accompanied by substantial shifts in configurational space.

Page 8: [67] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Like MD, NMR yields important insights, but

Page 9: [68] Deleted DECLAN CLARKE 7/28/15 12:09 AM

This type of investigation applied to many proteins simultaneously

Page 9: [68] Deleted DECLAN CLARKE 7/28/15 12:09 AM

This type of investigation applied to many proteins simultaneously

Page 9: [68] Deleted DECLAN CLARKE 7/28/15 12:09 AM

This type of investigation applied to many proteins simultaneously

Page 9: [69] Deleted DECLAN CLARKE 7/28/15 12:09 AM

data in

Page 9: [69] Deleted DECLAN CLARKE 7/28/15 12:09 AM

data in

Page 9: [69] Deleted DECLAN CLARKE 7/28/15 12:09 AM

data in

Page 9: [69] Deleted DECLAN CLARKE 7/28/15 12:09 AM

data in

Page 9: [70] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Thus, motivated by the idea that large differences in shape correspond to distinct conformations that occupy different energetic wells (Fig. 2), we describe and implement a pipeline for the identification of structures in distinct conformations using a statistical formalism. In doing so, we

Page 9: [70] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Thus, motivated by the idea that large differences in shape correspond to distinct conformations that occupy different energetic wells (Fig. 2), we describe and implement a pipeline for the identification of structures in distinct conformations using a statistical formalism. In doing so, we

Page 9: [71] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Our server readily enables the user to probe their own protein for potential allosteric regions, thereby helping to shed light on many

Page 9: [71] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Our server readily enables the user to probe their own protein for potential allosteric regions, thereby helping to shed light on many

Page 9: [71] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Our server readily enables the user to probe their own protein for potential allosteric regions, thereby helping to shed light on many

Page 9: [71] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Our server readily enables the user to probe their own protein for potential allosteric regions, thereby helping to shed light on many

Page 9: [71] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Our server readily enables the user to probe their own protein for potential allosteric regions, thereby helping to shed light on many

Page 9: [71] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Our server readily enables the user to probe their own protein for potential allosteric regions, thereby helping to shed light on many

Page 9: [71] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Our server readily enables the user to probe their own protein for potential allosteric regions, thereby helping to shed light on many

Page 9: [71] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Our server readily enables the user to probe their own protein for potential allosteric regions, thereby helping to shed light on many

Page 9: [71] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Our server readily enables the user to probe their own protein for potential allosteric regions, thereby helping to shed light on many

Page 9: [71] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Our server readily enables the user to probe their own protein for potential allosteric regions, thereby helping to shed light on many

Page 9: [71] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Our server readily enables the user to probe their own protein for potential allosteric regions, thereby helping to shed light on many

Page 9: [71] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Our server readily enables the user to probe their own protein for potential allosteric regions, thereby helping to shed light on many

Page 9: [71] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Our server readily enables the user to probe their own protein for potential allosteric regions, thereby helping to shed light on many

Page 9: [71] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Our server readily enables the user to probe their own protein for potential allosteric regions, thereby helping to shed light on many

Page 9: [71] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Our server readily enables the user to probe their own protein for potential allosteric regions, thereby helping to shed light on many

Page 9: [71] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Our server readily enables the user to probe their own protein for potential allosteric regions, thereby helping to shed light on many

Page 9: [71] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Our server readily enables the user to probe their own protein for potential allosteric regions, thereby helping to shed light on many

Page 9: [72] Deleted DECLAN CLARKE 7/28/15 12:09 AM

We employ a

Page 9: [72] Deleted DECLAN CLARKE 7/28/15 12:09 AM

We employ a

Page 9: [72] Deleted DECLAN CLARKE 7/28/15 12:09 AM

We employ a

Page 9: [72] Deleted DECLAN CLARKE 7/28/15 12:09 AM

We employ a

Page 9: [72] Deleted DECLAN CLARKE 7/28/15 12:09 AM

We employ a

Page 10: [73] Deleted DECLAN CLARKE 7/28/15 12:09 AM

finds

Page 10: [73] Deleted DECLAN CLARKE 7/28/15 12:09 AM

finds

Page 10: [73] Deleted DECLAN CLARKE 7/28/15 12:09 AM

finds

Page 10: [74] Deleted DECLAN CLARKE 7/28/15 12:09 AM

use a *mechanistic* approach for identifying

Page 10: [74] Deleted DECLAN CLARKE 7/28/15 12:09 AM

use a *mechanistic* approach for identifying

Page 10: [75] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Thus, we

Page 10: [75] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Thus, we

Page 10: [75] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Thus, we

Page 10: [75] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Thus, we

Page 10: [75] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Thus, we

Page 10: [75] Deleted DECLAN CLARKE 7/28/15 12:09 AM

Thus, we

Page 10: [76] Deleted DECLAN CLARKE 7/28/15 12:09 AM

any

Page 10: [76] Deleted DECLAN CLARKE 7/28/15 12:09 AM

any

Page 10: [76] Deleted DECLAN CLARKE 7/28/15 12:09 AM

any

Page 10: [76] Deleted DECLAN CLARKE 7/28/15 12:09 AM

any

Page 10: [76] Deleted DECLAN CLARKE 7/28/15 12:09 AM

any

Page 10: [76] Deleted DECLAN CLARKE 7/28/15 12:09 AM

any

Page 10: [76] Deleted DECLAN CLARKE 7/28/15 12:09 AM

any

Page 10: [76] Deleted DECLAN CLARKE 7/28/15 12:09 AM

any

Page 10: [77] Deleted DECLAN CLARKE 7/28/15 12:09 AM

That we achieved compelling results suggests that the level of coarse graining (i.e., in X-ray crystal structures and using ANMs instead of MD) was low enough to still recapitulate biologically interesting findings.

Page 10: [77] Deleted DECLAN CLARKE 7/28/15 12:09 AM

That we achieved compelling results suggests that the level of coarse graining (i.e., in X-ray crystal structures and using ANMs instead of MD) was low enough to still recapitulate biologically interesting findings.

Page 10: [77] Deleted DECLAN CLARKE 7/28/15 12:09 AM

That we achieved compelling results suggests that the level of coarse graining (i.e., in X-ray crystal structures and using ANMs instead of MD) was low enough to still recapitulate biologically interesting findings.

Page 10: [77] Deleted DECLAN CLARKE 7/28/15 12:09 AM

That we achieved compelling results suggests that the level of coarse graining (i.e., in X-ray crystal structures and using ANMs instead of MD) was low enough to still recapitulate biologically interesting findings.

Page 10: [77] Deleted DECLAN CLARKE 7/28/15 12:09 AM

That we achieved compelling results suggests that the level of coarse graining (i.e., in X-ray crystal structures and using ANMs instead of MD) was low enough to still recapitulate biologically interesting findings.

Page 10: [77] Deleted DECLAN CLARKE 7/28/15 12:09 AM

That we achieved compelling results suggests that the level of coarse graining (i.e., in X-ray crystal structures and using ANMs instead of MD) was low enough to still recapitulate biologically interesting findings.

Page 10: [77] Deleted DECLAN CLARKE 7/28/15 12:09 AM

That we achieved compelling results suggests that the level of coarse graining (i.e., in X-ray crystal structures and using ANMs instead of MD) was low enough to still recapitulate biologically interesting findings.

Page 12: [78] Deleted DECLAN CLARKE 7/28/15 12:09 AM

The K-values for MSAs, as well as the motivating conceptual framework, are summarized in Fig 2. About 3000 different domains had a K-value of 1 (i.e., one

conformation identified), whereas the K-values of close to 2000 domains exceed 1 (these exhibit multiple conformations, Fig. 2C). For proteins, close to 8000 had a K-value of 1, and about 1000 proteins had K-values that exceed 1. When performing K-means clustering with the gap statistic, very similar results were obtained when clustering structures on the basis of pairwise RMSD or pairwise Q_H (Supp. Fig. 3), so we use RMSD in our downstream analyses.

The fully-processed output for identifying high-confidence alternative conformations (which contains over 1100 proteins) is provided as a flat text file in the Supplementary content (Supp. File 1), and it is also included in our server as a downloadable text file. In addition to listing the alternative conformations, this file also lists descriptive statistics for each entry, including the RMSD between distinct conformers, cluster membership, degrees of confidence in assigning different structures to different clusters, etc.

We note that the pipeline above has been applied not only to SCOP domains, but also to individual proteins, with the only difference that only sequence-identical proteins were examined in this analysis.

Page 18: [79] Deleted **DECLAN CLARKE** **7/28/15 12:09 AM**

intra-protein communication).

After weights are assigned, the betweenness for each edge