SI METHODS

Identifying Potential Allosteric Residues

Identifying Surface-Critical Residues

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

We refer the reader to the work by Mitternacht and Berezovsky for details regarding the binding leverage method, though a general overview of the approach follows. Many candidate allosteric sites are generated by simulations in which a simple flexible ligand (comprising of 4 "atoms" linked by bonds of fixed length 3.8 Angstroms, but variable bond and dihedral angles) explores the protein's surface through many Monte Carlo steps. The number of MC simulations is set to 10 times the number of residues in the protein structure, and the number of MC steps within each simulation is set to 10,000 times the size of the simulation box, as measured in Angstroms. The size of this simulation box is set to 2x the maximum size of the PDB along any of the x, y or z-axes. *Apo* structures were used when probing protein surfaces for putative ligand binding sites in the canonical set of proteins.

A simple square well potential (i.e., modeling hard-sphere interactions) is used to model the attractive and repulsive energy terms associated with the ligand's interaction with the surface. These energy terms depend only on the ligand atoms' distance to alpha carbon atoms in the protein, and they are blind to other heavy atoms or biophysical properties. Once these candidate sites have been produced, normal mode analysis is applied is generate a model of the *apo* protein's low-frequency motions. Each of the candidate sites is then scored based on the degree to which deformations in the site couple to the low-frequency modes; that is, those sites which are heavily deformed as a result of the normal mode fluctuations receive a high score (termed the binding leverage for that site), whereas sites which undergo minimal change over the course of a mode fluctuation receive a low binding leverage score. The list of candidate sites is then processed to remove redundancy, and then ranked based on this score. 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 selective set of candidate sites. Indeed, the exclusion of other heavy atoms leaves 'holes' in the protein which do not actually exist in the context of the dense topology of side chain atoms. Thus, by including all heavy atoms, we hope to reduce the number of false positive candidate sites, as well as more realistically model ligand binding affinities in general.

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

Capturing Known Ligand-Binding Sites

Known ligand-binding residues are taken to be those within 4.5 Angstroms of the ligand within the *holo* structure (Supp. Table 1). 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. These statistics are achieved by covering an average of 15% of proteins' residues (Supp. Table 2), even though more than 15% of the proteins' residues are involved in ligand- or substrate-binding for most proteins (Supp. Table 3).

Dynamical Network Analysis to Identify Interior-Critical Residues

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

Specifically, the weight w_{ij} between residues i and j is set to $-\log(|C_{ij}|)$, where C_{ij} designates the correlated motions between residue i and j. If two contacting residues exhibit a high degree of correlated motion, then this implies that the motion of one residue may tell us about the motion of the other, suggesting a strong flow of energy or information between the two residues, resulting in a low value for w_{ij}. The 'network distance' between residues i and j (synonymous with w_{ij} in this discussion) is thus taken to be very short, and this short distance means that any path involving this pair of residues is shorter as a result, thereby more likely placing this pair of residues within any given shortest path, and more likely rendering this pair of residues a bottleneck pair. In sum, a high correlation in motion results in a short distance, thereby more likely rendering this a bottleneck pair of residues.

Finally, once all connections between contacting pairs are appropriately weighted and the communities are assigned, a residue is deemed to be critical for allosteric signal transmission (i.e., an interior-critical residue) 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 interior-critical residues highlighted in Supp. Figs. 3 and 4.

Decomposing Proteins into Modules Using Different Algorithms

Many algorithms have been devised to extract the community structure of networks. In a comprehensive study comparing different algorithms (Lancichinetti et al, 2009), an information theory-based approach (Rosvall et al, 2007), was shown to be one of the strongest. This method (termed "Infomap") effectively reduces the network community detection problem to a problem in information compression: the prominent features of the network are extracted in this compression process, giving rise to distinct modules (more details are provided in Rosvall et al, 2007).

Perhaps surprisingly, even though both Infomap and GN achieve similar network modularity, we find that Infomap (see Methods and Rosvall et al, 2007) produces at least twice the number of communities relative to that of GN, and it thus generates many more interior-critical residues (Supp. Table 6 and Supp. Fig. 20). For the canonical set of proteins, GN and Infomap generated an average of 12.0 and 36.8 communities, respectively (corresponding to an average of 44.8 and 201.4 interior-critical residues, respectively). Thus, given that GN produces a more selective set of residues for each protein, the focus of our analyses is based on GN.

Although the critical residues identified by GN do not always correspond to those identified by Infomap, the mean fraction of GN-identified interior-critical residues that match Infomap-identified residues is 0.30 (the expected mean is 0.21, p-value=0.058), which further justifies our decision to focus on GN). Furthermore, we observe that obvious structural communities are detected when applying both methods (i.e., a community generated by GN is often the same as that generated by Infomap, and in other cases, a community generated by GN is often composed of sub-communities generated by Infomap).

As noted, the modularity from the network partitions generated by GN and Infomap are very similar (for the 12 canonical systems, the mean modularity for GN and Infomap is 0.73 and 0.68, respectively). Presumably, GN modularity values are consistently at least as high as those in Infomap because GN explicitly optimizes modularity in partitioning the network, whereas Infomap does not.

STRESS (STRucturally-identified ESSential residues)

Our server has been designed to be both user-friendly and highly efficient. We use locality-sensitive hashing to do local search in each sampling step in the search for surface-critical residues, which takes constant time. The time complexity of the core computation, Monte Carlo sampling, is O(|T||S|), where T and S are simulation trials and steps for each trial, respectively. After carefully profiling and optimization, a typical case takes only about 30 minutes on one E5-2660 v2 (2.20GHz) core.

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

High-Throughput Identification of Alternative Conformations

An overview of our pipeline is provided in Supp. Fig. 9, 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 use information regarding protein motions to identify potential allosteric sites on the surface and within the interior.

Database-Wide Multiple Structure Alignments

FASTA files of all SCOP domains were downloaded from the SCOP website (version 2.03) (Murzin et al, 1995; Fox et al, 2014). We first worked with domains to probe for intra-domain conformational changes, as better structure alignments are generally possible at the domain level.

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 (Altschul et al, 1997) 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 (Kosloff et al, 2008), 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 (as of December 2013).

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 (Russell et al, 1992) and MultiSeq (Roberts et al, 2006) plugins of VMD (Humphrey et al, 1996) 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) 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. Q_H, an alternative metric to RMSD, quantifies the degree to which residue-residue distances differ between two conformations, and is detailed in (O'Donoghue et al, 2003). 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. PDB-wide MSAs across sequence-similar groups reveal that, in agreement with expectation, average pairwise root-mean-square deviation (RMSD) values increase at lower levels of sequence identity, as do Q_H values) (Supp. Fig. 21).

Identifying Distinct Conformations within a Multiple Structure Alignment

For each MSA produced in the previous step (using only sets of sequences that are 100% sequence-identical), 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 structure, 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 identifying the number of distinct conformational states in an ensemble of structures relies on a modified version of the K-means clustering algorithm. This modified form of the algorithm is termed K-means clustering with the gap statistic, and it was introduced in (Tibshirani et al, 2001).

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 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 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 (Gower et al, 1996; Mardia et al, 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 structures 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, as well as more details on the theoretical justifications of this approach. However, an overview of the general intuition behind this formalism is provided here.

For the purpose of demonstration, assume that the data takes the form of 60 data points, with each point represented in 2D space (in variables x and y). See blue points Supp. Fig. 22. Of course, our observed data in the case of multiple structure alignments may lie in N-dimensional space, in which case all Euclidean distances are just as easily calculated.

1) Assume that the input data can be represented with K clusters. Perform Lloyd's algorithm on the dataset in order to assign each point to one of the K clusters. Then, for each cluster k, measure D_k, which describes the ‘density’ of points within cluster k:

$$D_k = \sum_{\mathrm{x}_i \in C_k} \sum_{\mathrm{x}_j \in C_k} ||\mathrm{x}_i - \mathrm{x}_j||^2$$

2) Calculate an overall normalized score W to describe how well-clustered the resultant system has become when assigning all 60 data points to the K clusters:

$$W \ = \sum_{k=1}^{K} \frac{1}{2n_k} D_k$$

3) Given our observed data, how well does this number of assigned clusters actual represent the ‘true’ number of clusters represented by the data, relative to a null model without any clustering? To address this question, generate a null model by producing 60 randomly-distributed data points that lack any clear clustering (grey points in Supp. Fig. 22) such that the randomly-placed points lie within the same bounding box of the observed data (in blue).

4) Repeat step (3) above M times, and each time a random null distribution is produced, calculate $W_{null(K)}$ for each distribution (assuming K clusters), just as W is calculated for the observed data. Then calculate the mean$_{\mathrm{M}}\{\log(W_{null(K)})\}$ for these M null

distributions. Intuitively, the value $\text{mean}_M\{\log(W_{null(K)})\}$ measures how well *random* systems (with the same number of data points and within the same variable ranges as the observed data) can be described by K clusters. The M $\log(W_{null(K)})$ values produced by the null models have a standard deviation that is ultimately converted to the following (see Tibshirani et al, 2001 for details):

$$s_k = \sqrt{1 + 1/B}\,\text{sd}(k)$$

5) Calculate the gap statistic δ(K), given K clusters. Intuitively, a high value for this statistic signifies that our data is well-described using K clusters, relative to the assignment of K clusters in a randomized null distribution. Assuming K clusters, the gap statistic is given as:

$$\delta(K) = \text{mean}_M\{\log(W_{null(K)})\} - \log(W)$$

6) Obtain the values δ(K+1), δ(K+2), δ(K+3), etc. This is done simply by incrementing the value for K and repeating the steps (1) through (5) above. Note that the optimal value of K ($K_{optimal}$, which is 3 in our demonstration case) is taken to be the first (i.e., lowest) K such that $\delta(K) >= \delta(K+1) - s_{k+1}$:

$$K_{optimal} = \{K |\ \delta(K) >= \delta(K+1) - s_{k+1}\}$$

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 (Rodgers et al, 2013; Swain et al, 2006)).

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 and protein chains.

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 (Murtagh et al 1985), with UPGMA (mean values) used as the chosen agglomeration method (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).

Modeling Protein Motions by Directly Using Displacement Vectors from Alternative Conformations

As discussed, conformational changes may be modeled using vectors connecting pairs of corresponding residues in crystal structures of alternative conformations (termed "ACT"). This more direct model of conformational change is especially straightforward to apply to single-chain proteins (applying this method on a database scale to multi-chain complexes introduces confounding factors related to chain-chain correspondence between such complexes when each complex has multiple copies of a given chain).

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. 13, left), and the same trend is observed when this is applied in our dynamical network analysis for identifying interior-critical residues (Supp. Fig. 13, right). There are too few human single-chain proteins to perform a reliable analysis in which conservation is evaluated using 1000 Genomes or ExAC data – for instance, only 9 (16) structures are such that 1000 Genomes SNVs (ExAC SNVs) overlap with interior-critical residues.

Evaluating the Conservation of Critical Residues with Various Metrics and Data Sources

Conservation Across Species

All cross-species conservation scores represent the ConSurf scores, as downloaded from the ConSurf Server (Ashkenazy et al, 2010; Celniker et al, 2013; Glaser et al, 2003; Landau et al, 2005), in which scores for each protein chain are normalized to 0. Low (i.e., negative) ConSurf scores represent a stronger degree of conservation, and high (i.e., positive) scores designate weaker conservation. We perform cross-species conservation analysis on those proteins for which ConSurf files are available from the ConSurf server, and all ConSurf scores were calculated using default parameters, listed here:

Homolog search algorithm: CSI-BLAST
Number of iterations: 3
E-value cutoff: 0.0001
Proteins database: UniRef-90
Maximum homologs to collect: 150
Maximal %ID between sequences: 95
Minimal %ID for homologs: 35
Alignment method: MAFT-L-INS-i
Calculation method: Baysian
Calculation method: JTT

Each individual point within the cross-species conservation plots (e.g., Figs. 3B and 3F, and Supp. Fig. 13) represents data from one protein: the value of the point for any given protein represents the mean conservation score for all residues within one of two classes: the set of N critical residues within a protein structure (surface or interior) or a randomly-selected set of N non-critical residues (with the same "degree", see proceeding paragraph below) 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 distributions of non-critical residues shown are very much representative of the distributions observed when re-building the random set many times.

Note that the degree (i.e., k) of residue j is defined as the number of residues which interact with residue j, where residues adjacent to residue j in sequence are not considered, and an interaction is defined whenever any heavy atom in an interacting residue is within 4.5 Angstroms of any heavy atom in residue j. We use degree as a measure of residue burial for several reasons. Our use of degree as a metric for characterizing burial is consistent with our networks-based analysis for identifying interior-critical residues, as well as our use of residue-residue contacts in building networks for producing the ANMs. Residue degree is also an attractive metric because it is discrete in nature, thereby allowing us to generate null distributions of non-critical residues with the exact same degree distribution.

Measures of Conservation Amongst Humans from Next-Generation Sequencing

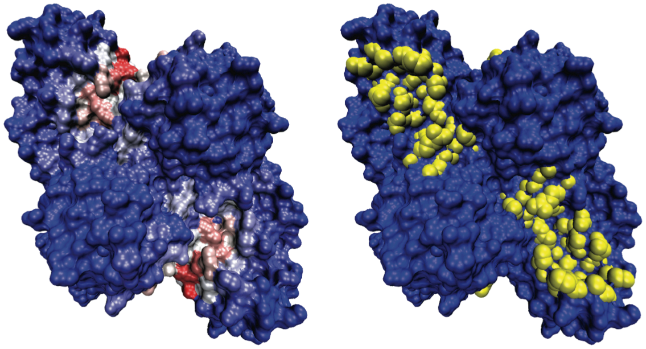
All SNVs hitting protein-coding regions that result in amino acids changes (i.e., nonsynonymous SNVs) were collected from The 1000 Genomes Project (phase 3 release) (1000 Genomes Project Consortium, 2012). VCF files containing the annotated variants were generated using VAT (Habegger et al, 2012). For 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 (Smedley et al, 2015). 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).

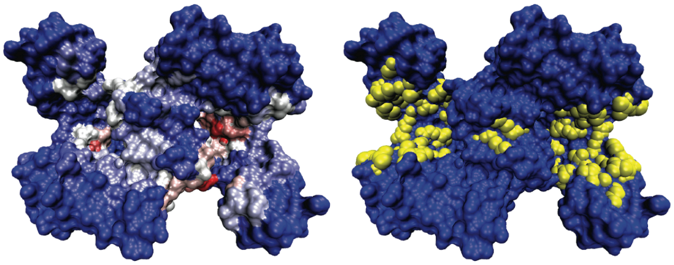
Only structures for which at least one critical residue and one non-critical residue are hit by ExAC SNVs are included in the analysis. As with the 1000 Genomes analysis, this enables a more direct comparison between critical and non-critical residues, as comparisons between two different proteins would rely on the assumption of equal degrees of selection between such proteins.

Each individual point within the intra-human conservation plots (e.g., Figs. 3C, 3D, 3G, 3H) represents data from one protein: the value of the point for any given protein represents the mean score (DAF or MAF, for 1000 Genomes or ExAC variants, respectively) for all critical (red bars) or non-critical (blue bars) residues to be hit by SNVs.

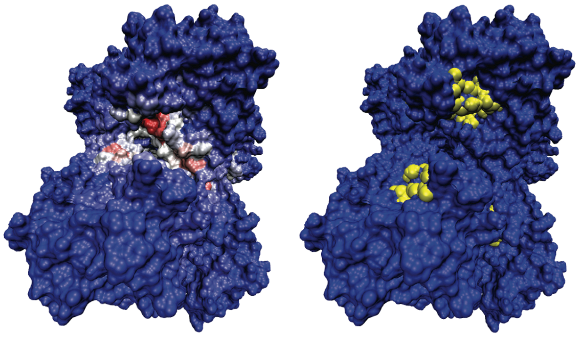
Supp. Fig. 1: Set of canonical proteins. Left images designate sites that are scored highly (i.e., surface-critical residues), and right images show the residues (yellow) that actually come into contact of known ligands.



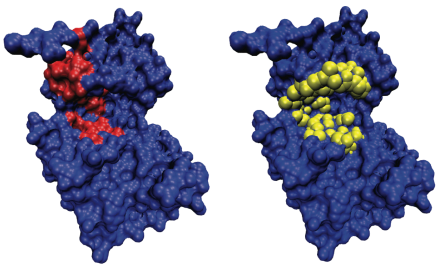
3pfk Phosphofructokinase



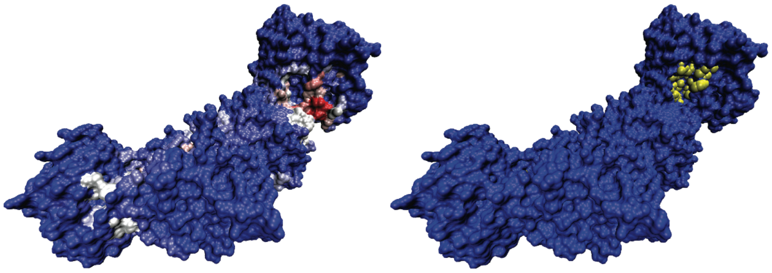
4ake Adenylate Kinase



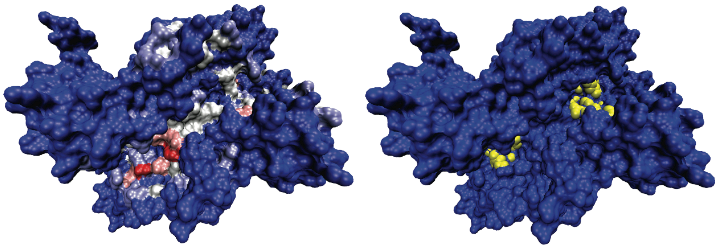
1cd5 G6P-Deaminase



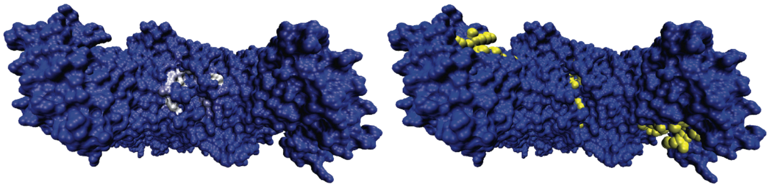
1j3h cAMP-dependent Kinase



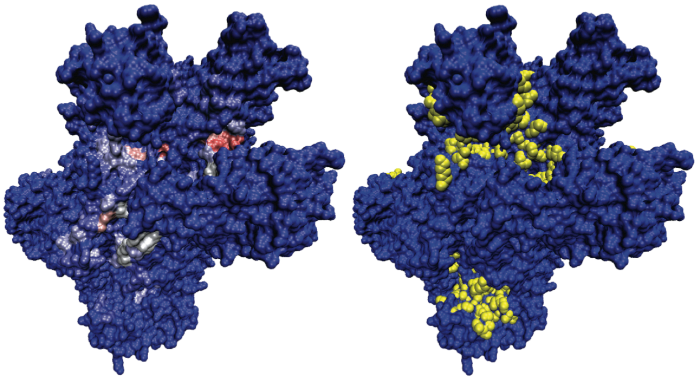
1bks Trp Synthase



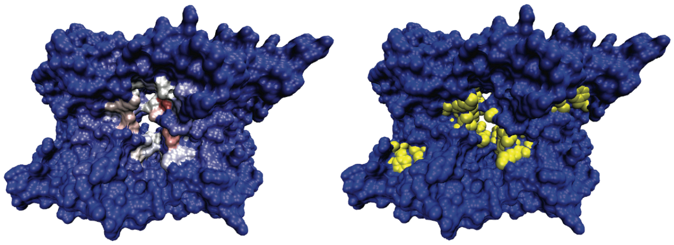
1e5x Thr Synthase



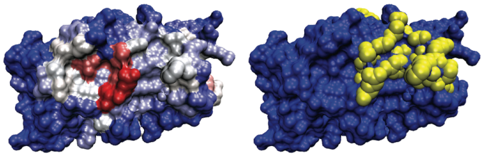
1efk Malic Enzyme



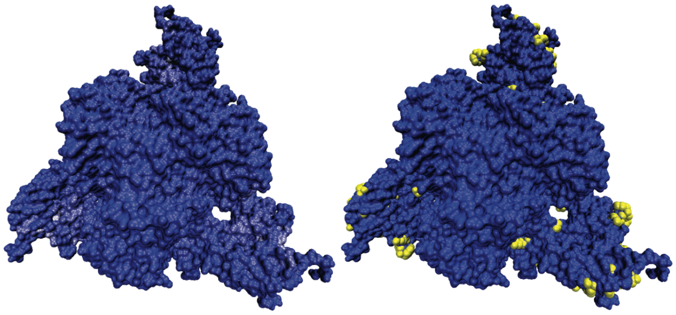
1nr7 Glu Dehydrogenase



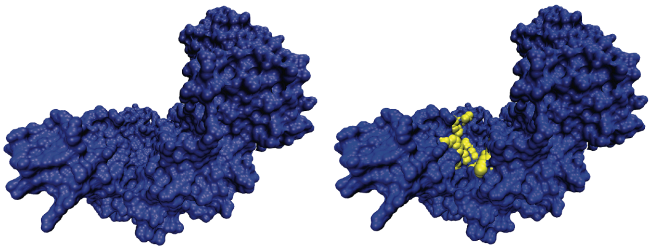
1xtt Phosphoribosyltransferase



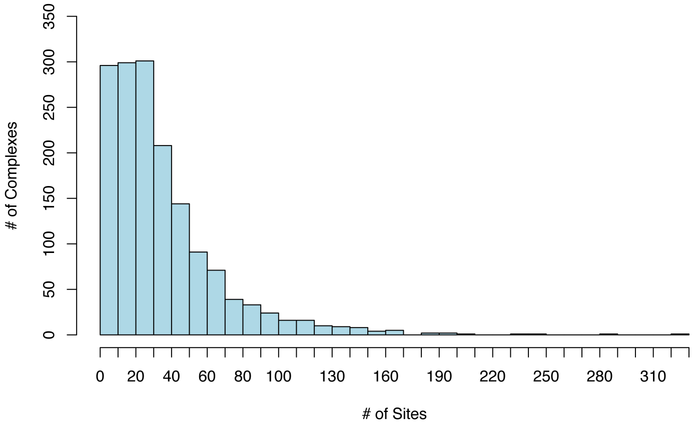
2hnp Tyr Phosphatase



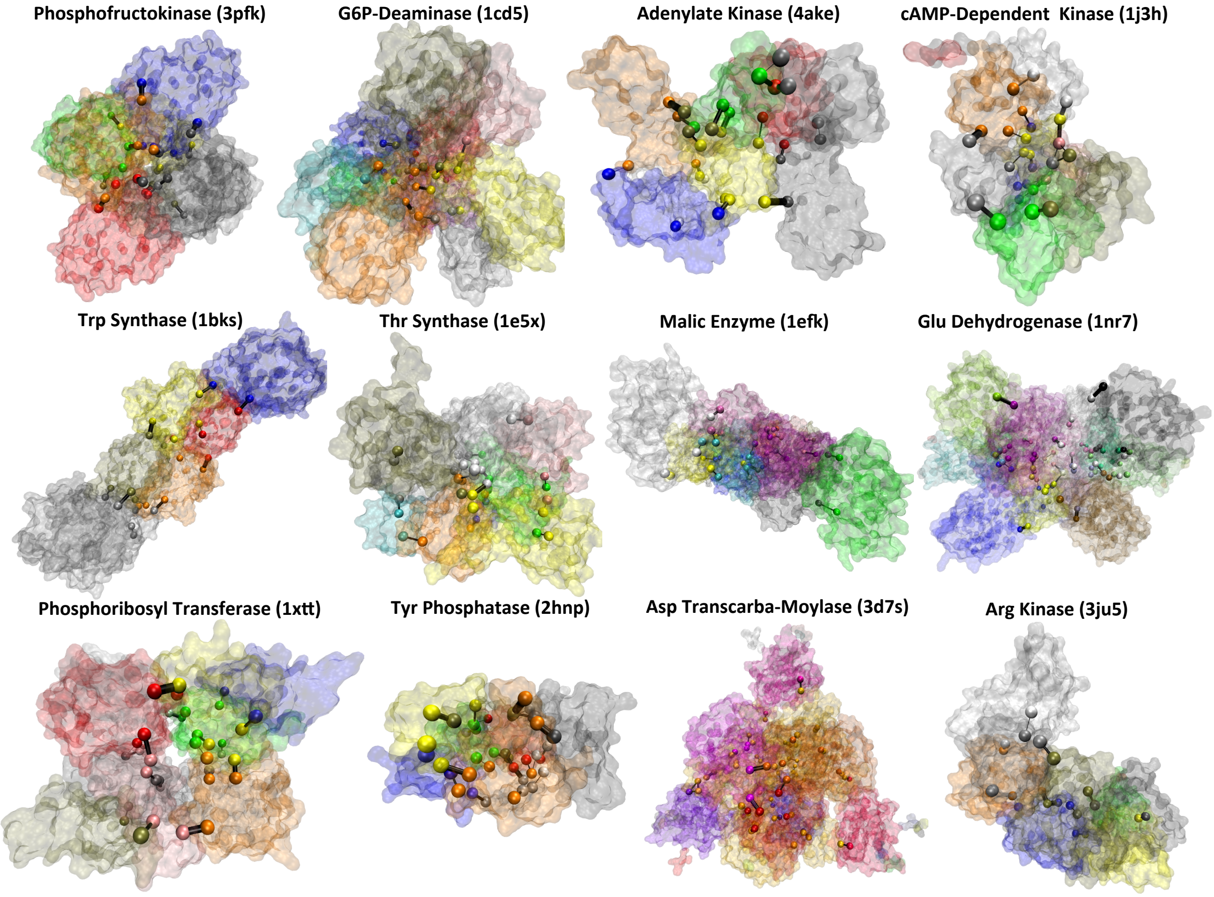
3d7s Asp Transcarbamoylase



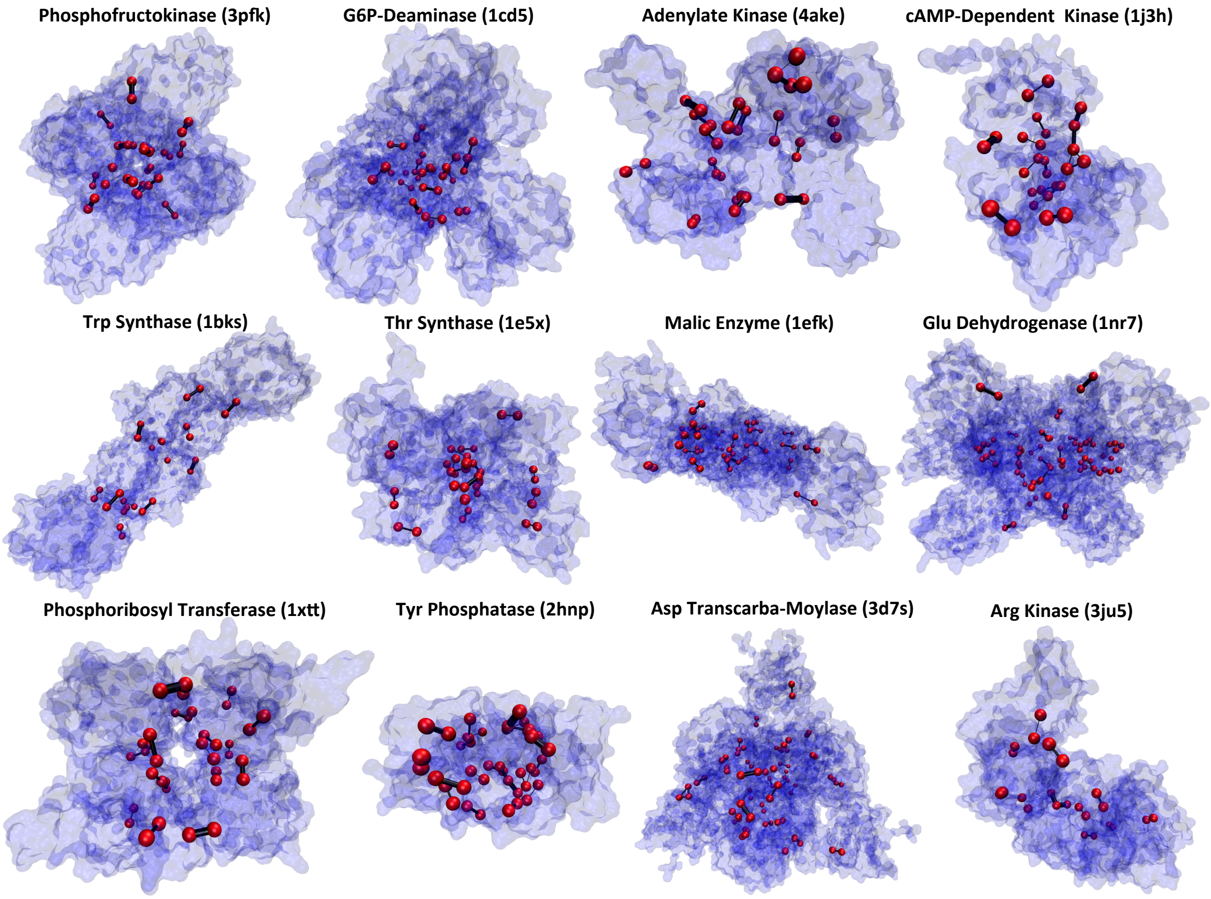
3ju5 Arg Kinase



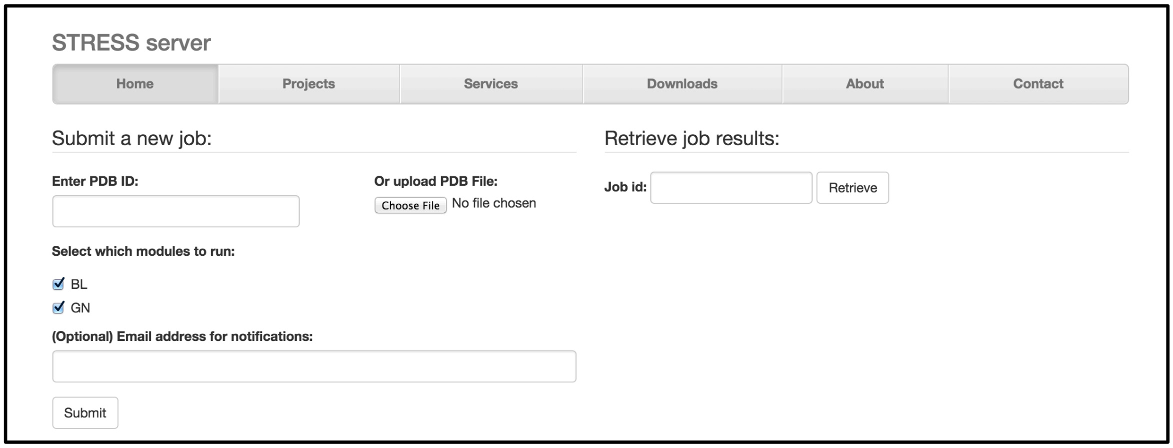
Supp Fig. 2: Number of surface-critical sites per complex without thresholding



Supp Fig. 3: Community partitioning for example systems
Communities identified by dynamical network-based analysis. Different communities are colored differently. Residues shown as spheres are interior-critical residues. The thickness of a black links between a pair of residues is proportional to that pair's associated betweenness.



Supp Fig. 4: Interior-critical residues highlighted in several example systems. The same structures as those given in Supp. Fig. 3, but with interior-critical residues highlighted in red spheres.



Submit a new job:

Enter PDB ID:

Or upload PDB File: No file chosen

Select which modules to run:

☑ BL

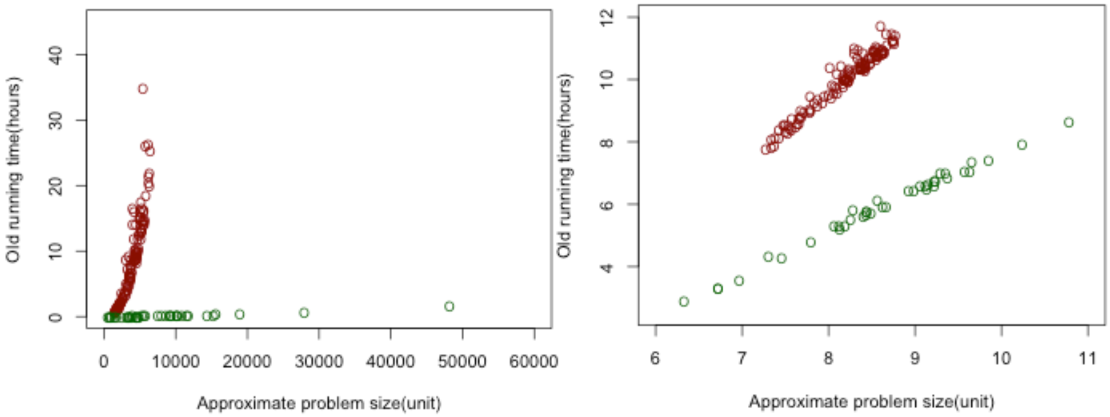
☑ GN

(Optional) Email address for notifications:

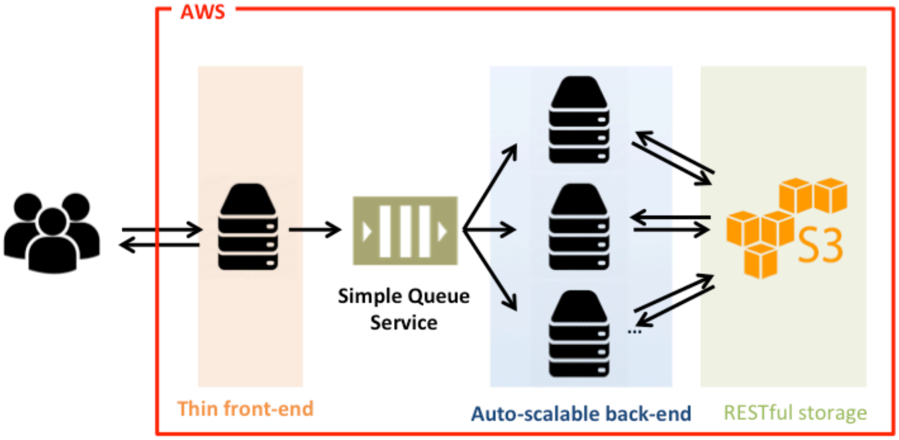
Retrieve job results:

Job id:

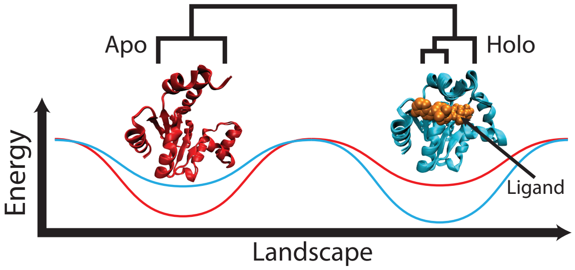
Supp. Fig. 5: Main page of STRESS server (stress.gersteinlab.org)



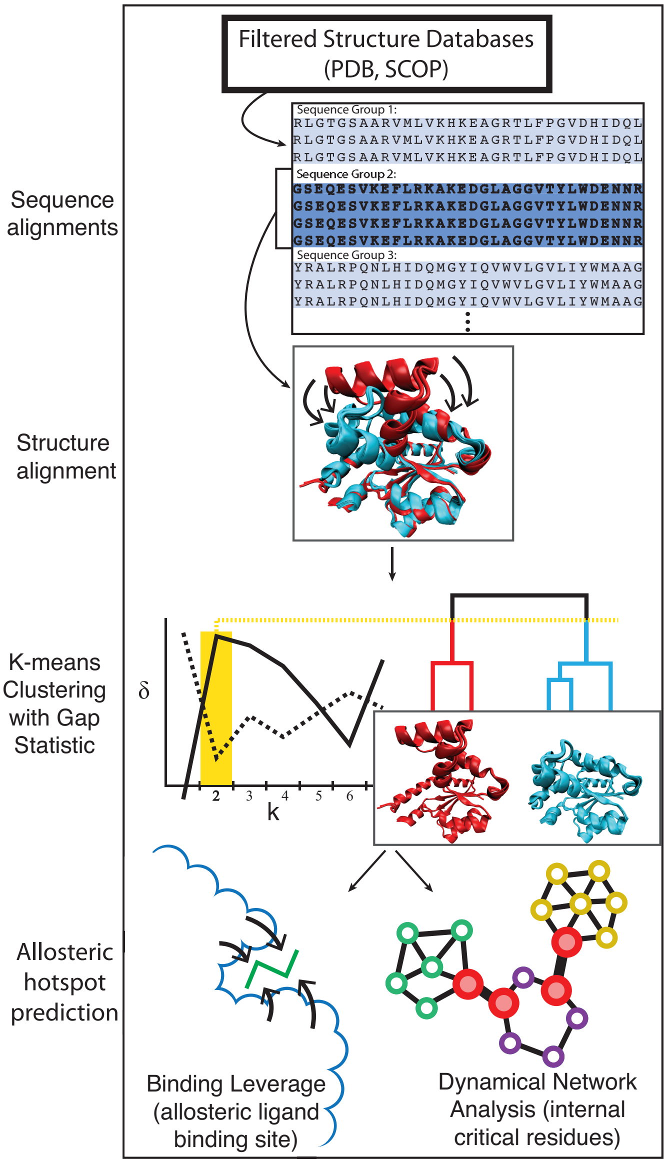
Supp. Fig. 6: Code optimization in the search of surface-critical residues. Running times are shown for systems of various sizes.



Supp. Fig. 7: Architecture of STRESS server

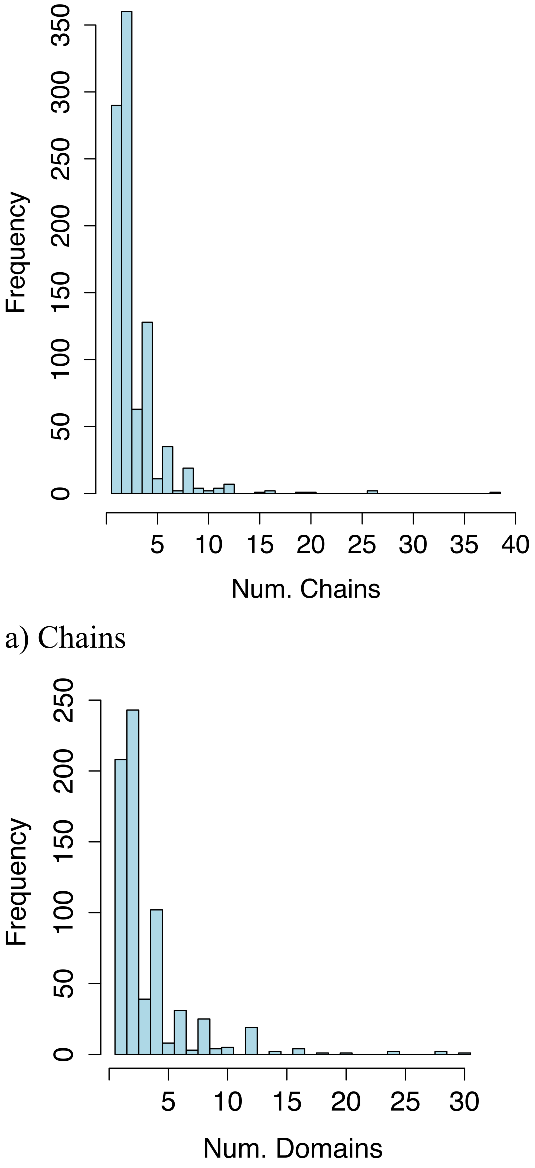


Supp. Fig. 8


Filtered Structure Databases
(PDB, SCOP)
Sequence alignments
Sequence Group 1:
RLGTGSAARVMLVKHKEAGRTLFPGVDHIDQL
RLGTGSAARVMLVKHKEAGRTLFPGVDHIDQL
RLGTGSAARVMLVKHKEAGRTLFPGVDHIDQL
Sequence Group 2:
GSEQESVKEFLRKAKEDGLAGGVTYLWDENNR
GSEQESVKEFLRKAKEDGLAGGVTYLWDENNR
GSEQESVKEFLRKAKEDGLAGGVTYLWDENNR
GSEQESVKEFLRKAKEDGLAGGVTYLWDENNR
Sequence Group 3:
YRALRPQNLHIDQMGYIQVWVLGVLIYWMAAG
YRALRPQNLHIDQMGYIQVWVLGVLIYWMAAG
YRALRPQNLHIDQMGYIQVWVLGVLIYWMAAG
Structure alignment
K-means Clustering with Gap Statistic
δ
2
3
4
5
6
k
Allosteric hotspot prediction
Binding Leverage (allosteric ligand binding site)
Dynamical Network Analysis (internal critical residues)

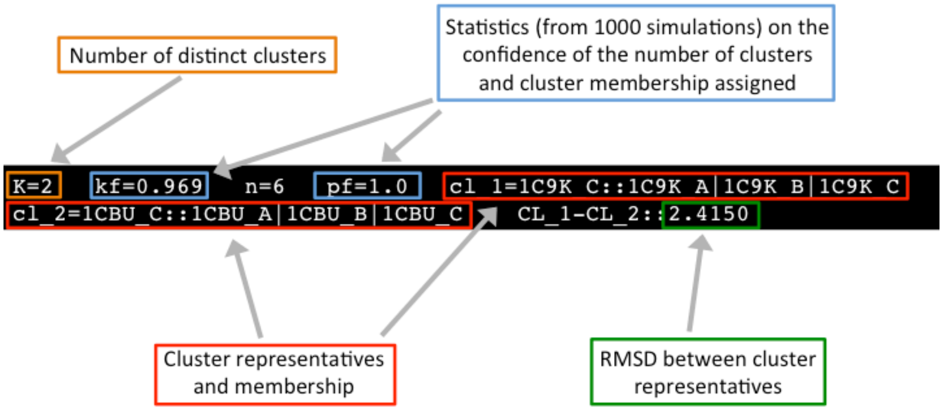
Supp Fig. 9: Pipeline for identifying distinct conformations and critical residues *Top to bottom*: **a)** BLAST-CLUST is applied to the sequences corresponding to a filtered set of structures, thereby providing a large number of sequence-identical groups. **b)** For each sequence-identical group, a multiple structure alignment 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 structures which exhibit multiple clusters (i.e., those with K > 1) are then taken to exhibit multiple conformations.

Supp Fig. 10: Distributions of the number of chains and domains in set of alternative conformations



a) Chains

b) Domains

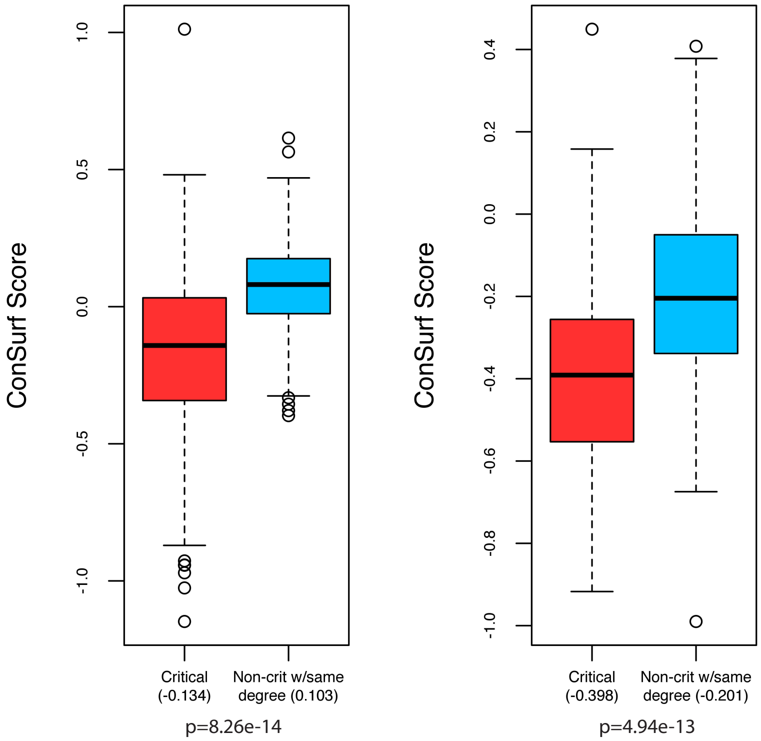


```
K=2   kf=0.969   n=6   pf=1.0   cl_1=1C9K_C::1C9K_A|1C9K_B|1C9K_C
cl_2=1CBU_C::1CBU_A|1CBU_B|1CBU_C    CL_1-CL_2::2.4150
```

Number of distinct clusters
Statistics (from 1000 simulations) on the confidence of the number of clusters and cluster membership assigned
Cluster representatives and membership
RMSD between cluster representatives

Supp. Fig. 12: A single annotated entry from our database of alternative conformations.

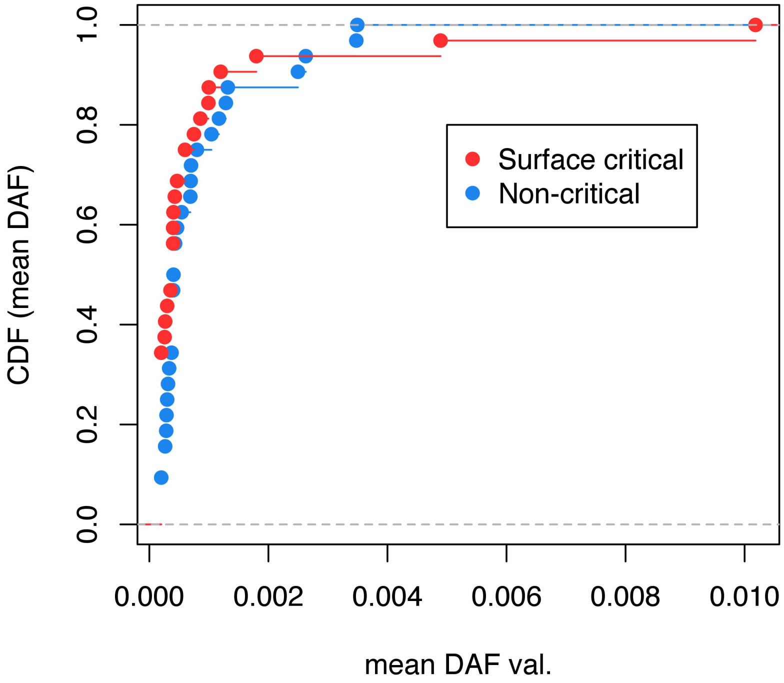
The clustering for the protein adenosylcobinamide kinase is shown. 2 distinct conformations are represented in the ensemble of structures. The measure *kf* designates the fraction of times that the optimal value of K (here, K=2) was obtained out of 1000 simulations in which the algorithm (K-means with the gap statistic) obtained this particular value of K. The high *kf* value (0.969) signifies that these structures are very well clustered into two groups. *n* designates the number of distinct structures (PDB chains in this case) in the multiple structure alignment. *pf* designates the fraction of times (out of 1000 simulations of running Lloyd's algorithm, the standard K-means algorithm) that this particular set of structure-group assignments were assigned. In this this example, for all 1000 simulations, 1C9K_C and 1C9K_A were clustered in one group, and 1CBU_A, 1CBU_B, 1CBU_C clustered together. Within each cluster (the two clusters shown as two red boxes), the chain preceding the "::" tag designates the cluster representative (i.e., the structure closest to the Euclidean centroid of the cluster). The last field gives the RMSD values between cluster representatives. See the header information within Supp. File 1 for further details.



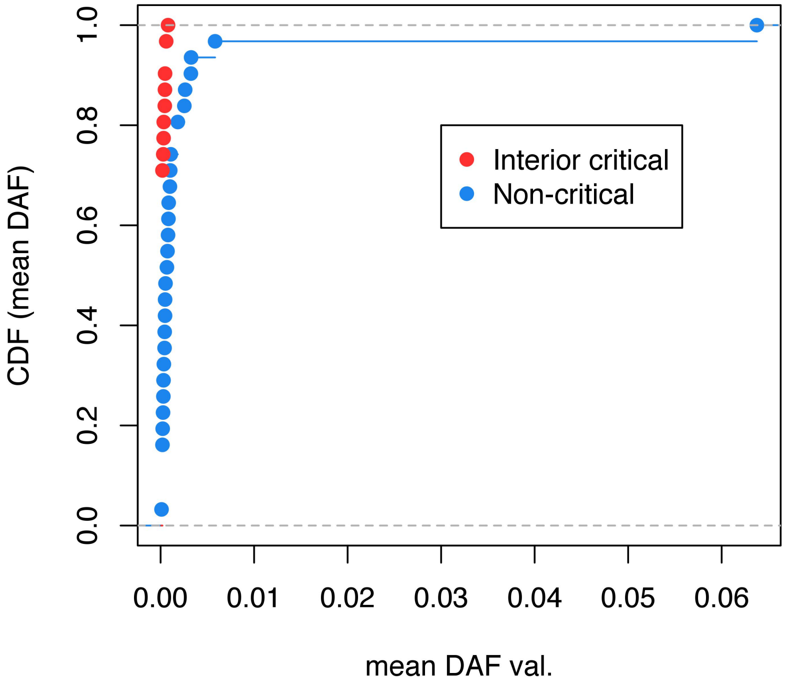
Supp Fig. 13: Modeling protein conformational change through a direct use of crystal structures from alternative conformations using absolute conformational transitions (ACT)

Left: Distributions of the mean conservation scores on surface-critical (red) and non-critical residues with the same degree of burial (blue). *Right*: Distributions of the mean conservation scores for interior-critical (red) and non-critical residues with the same degree of burial (blue). Mean values are given in parentheses. Results for single-chain proteins are shown, and p-values were calculated using a Wilcoxon rank sum test.

Supp. Fig. 14: Potential shifts in DAF distributions (in 1000 Genomes) using two-sample Kolmogorov-Smirnov tests

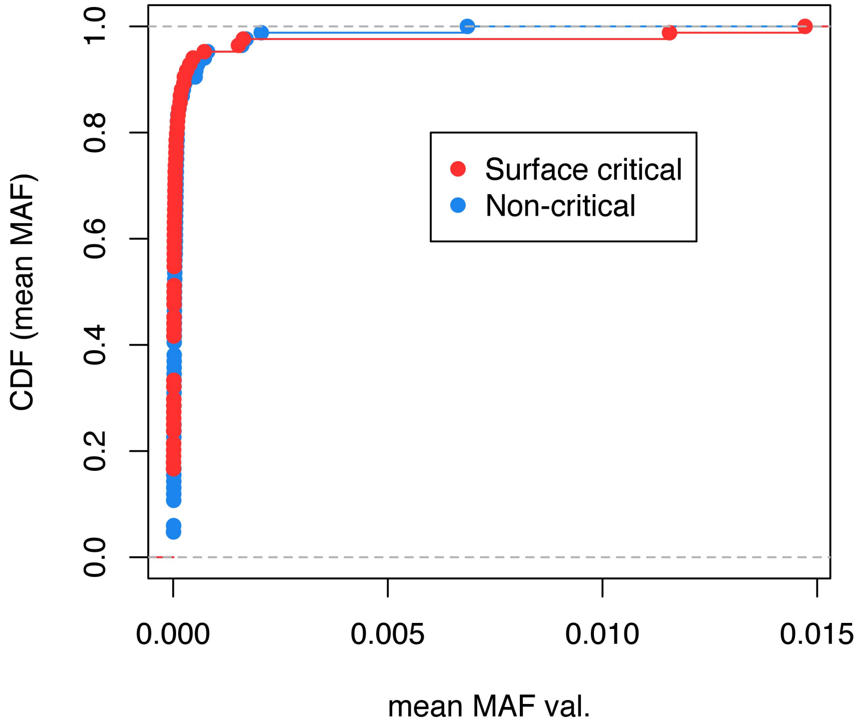


14 A) Cumulative distribution functions for mean DAF values of surface-critical and non-critical residues (p-val = 0.159).

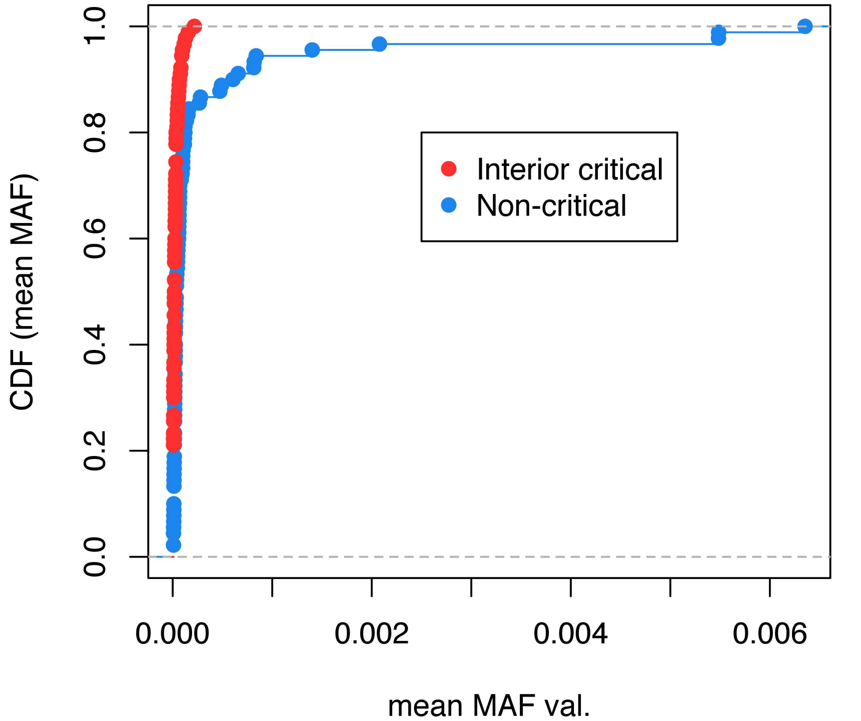


14 B) Cumulative distribution functions for mean DAF values of interior-critical and non-critical residues (p-val = 1.79e-4).

Supp Fig. 15: Potential shifts in mean minor allele frequency distributions (in ExAC) using two-sample Kolmogorov-Smirnov tests



15 A) Cumulative distribution functions for mean minor allele frequencies of surface-critical and non-critical residues (p-val = 9.49e-2).



15 B) Cumulative distribution functions for mean minor allele frequencies of interior-critical and non-critical residues (p-val = 1.75e-4).

pdb	Fract rare SNPs in CRIT	Fract rare SNPs in NON crit
2F0Y	1	0.470588235
3GLS	0.5	0.875
1I3L	1	0.9
1T09	1	0.912234043
1GG3	1	0.9375
1T0L	1	0.951612903
1DE4	1	0.958333333
1BX4	1	1
1H6G	1	1
1HZD	1	1
1IIL	1	1
1MMK	1	1
1XRJ	1	1
1ZNQ	1	1
1ZVM	1	1
2AH9	1	1
2FY7	1	1
2O3T	1	1
2ONM	1	1
2ZQQ	1	1
3B6R	1	1
3BL7	1	1
3DRB	1	1
3FVX	1	1
3I7G	1	1
3KEJ	1	1
3KMW	1	1
3RPN	1	1
3RPP	1	1
3ZNS	0	1
4F45	1	1
4H9S	1	1

PDB	Fract rare SNPs in CRIT	Fract rare SNPs in NON-CRIT
2F0Y	1	0.470588235
2O3T	1	0.535714286
4F45	1	0.694117647
1I3L	1	0.8
1ZVM	1	0.820689655
1HZD	1	0.833333333
1ZNQ	1	0.833333333
2ZQQ	1	0.833333333
3B6R	1	0.857142857
4H9S	1	0.857142857
1GG3	0.75	0.875
3GLS	0.5	0.875
3I7G	0.516129032	0.884057971
3DRB	1	0.888888889
3KEJ	0	0.896103896
1T09	0.806451613	0.912234043
1DE4	1	0.916666667
3RPP	1	0.916666667
3RPN	1	0.939393939
3BL7	1	0.944444444
1T0L	1	0.951612903
1BX4	1	1
1H6G	1	1
1IIL	1	1
1MMK	1	1
1XRJ	1	1
2AH9	1	1
2FY7	1	1
2ONM	0	1
3FVX	0.959459459	1
3KMW	1	1
3ZNS	0	1

Supp. Fig. 16: Defining the fraction of rare variants using 1000 Genomes data for surface-critical residues

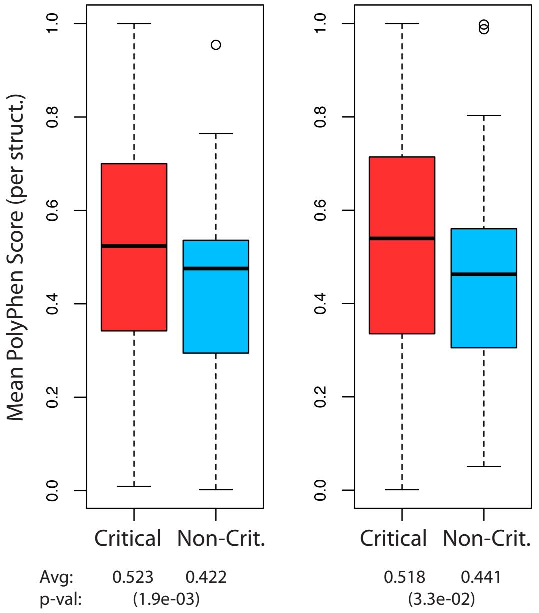
Fraction of rare 1000 Genomes alleles (using a DAF cutoff of 0.05% and 0.01%, for the left and right lists, respectively) for surface-critical and non-critical residues. Green is used to highlight cases for which the fraction of rare variants is higher in surface-critical residues than in non-critical residues, and gray designates cases for which the opposite trend is observed.

pdb	Fract rare SNPs (crit)	Fract rare SNPs (NON crit)
2WP3	1	0.666666667
1LD7	1	0.742857143
2F0Y	1	0.742857143
3GLS	1	0.777777778
3C10	1	0.785714286
1JDX	1	0.833333333
2R1V	1	0.920792079
1S1P	1	0.924882629
1T0L	1	0.951612903
1DE4	1	0.961538462
1GG3	1	1
1H6G	1	1
1IIL	1	1
1MMK	1	1
1RKB	1	1
1W24	1	1
1ZVM	1	1
2AH9	1	1
2OO1	1	1
3EVX	1	1
3FVX	1	1
3HPH	1	1
3I7G	1	1
3KEJ	1	1
3KMW	1	1
3LJZ	1	1
3O5M	1	1
3RPN	1	1
3RPP	1	1
4F45	1	1
4HW3	1	1

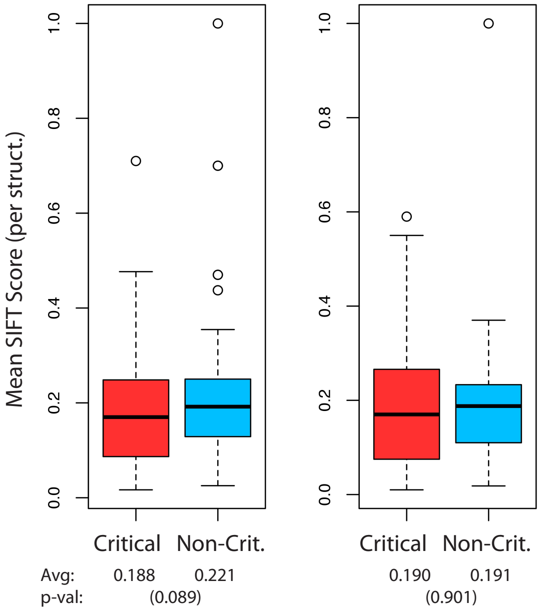
PDB	Fract rare in CRIT	Fract rare SNPs in NON-CRIT
2WP3	1	0.666666667
3O5M	1	0.73015873
1LD7	1	0.742857143
2F0Y	1	0.742857143
3LJZ	1	0.75
1ZVM	1	0.771929825
3GLS	1	0.777777778
1S1P	1	0.784037559
3C10	1	0.785714286
3I7G	1	0.797385621
3KEJ	1	0.798701299
2R1V	1	0.811881188
1GG3	1	0.818181818
1JDX	1	0.833333333
1DE4	1	0.846153846
4HW3	1	0.846153846
3RPP	1	0.923076923
3RPN	1	0.9375
1T0L	1	0.951612903
3FVX	1	0.966292135
1H6G	1	1
1IIL	1	1
1MMK	1	1
1RKB	1	1
1W24	1	1
2AH9	1	1
2OO1	1	1
3EVX	1	1
3HPH	1	1
3KMW	1	1
4F45	0.821918	1

Supp. Fig. 17: Defining the fraction of rare variants using 1000 Genomes data for interior-critical residues

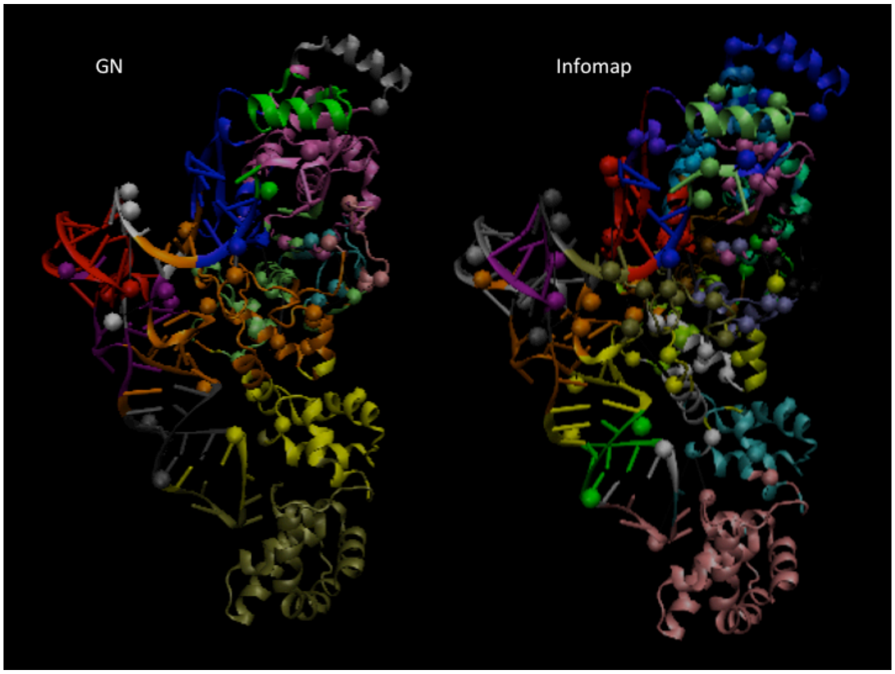
Fraction of rare 1000 Genomes alleles (using a DAF cutoff of 0.05% and 0.01% in the left and right lists, respectively) for interior-critical and non-critical residues. Green is used to highlight cases for which the fraction of rare variants is higher in interior-critical residues than in non-critical residues.



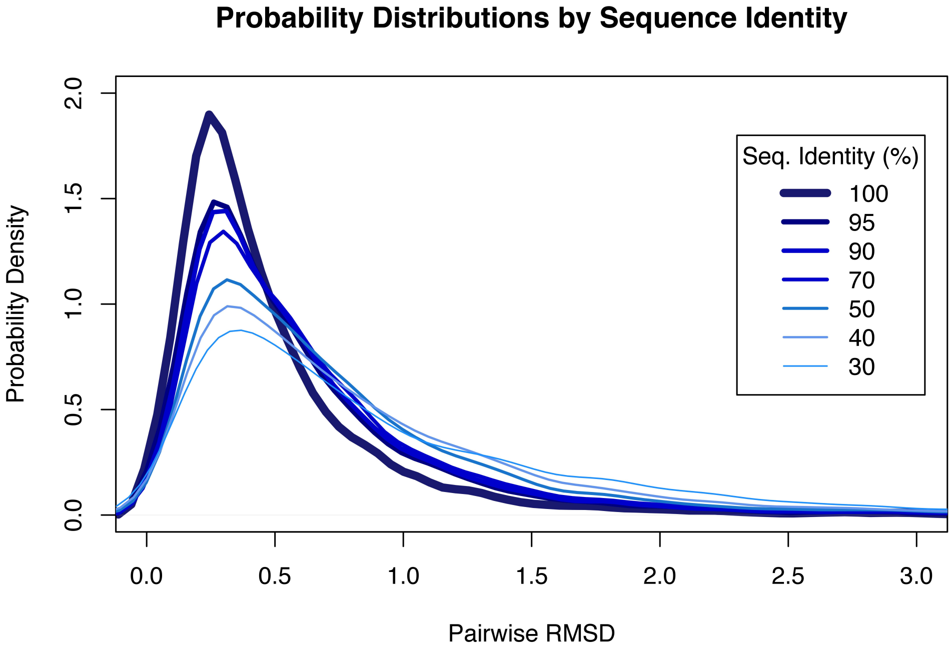
Supp Fig. 18: Mean PolyPhen scores for critical- and non-critical residues, as identified by ExAC. *Left*: Distribution of mean PolyPhen values on surface-critical residues (red) and non-critical residues (blue). *Right*: Distribution of mean PolyPhen values on interior-critical residues (red) and non-critical residues (blue). Overall mean values and p-values are given below plots. Note that higher PolyPhen scores denote more damaging variants.



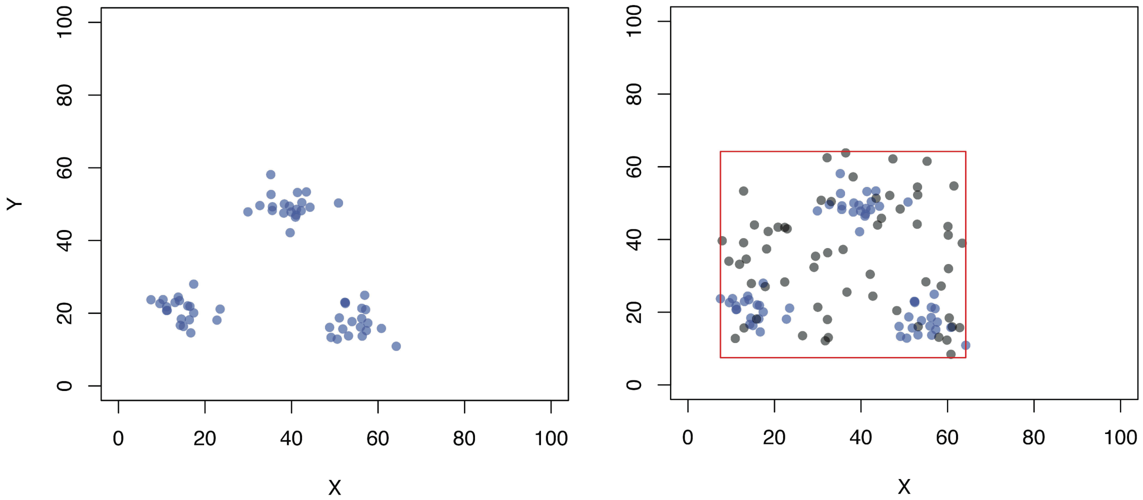
Supp Fig. 19: Mean SIFT scores for critical- and non-critical residues, as identified by ExAC. *Left*: Distribution of mean SIFT values on surface-critical residues (red) and non-critical residues (blue). *Right*: Distribution of mean SIFT values on interior-critical residues (red) and non-critical residues (blue). Overall mean values and p-values are given below plots. Note that lower SIFT scores denote more damaging variants.



Supp. Fig. 20: Network modularization by GN and Infomap
Different colors correspond to different communities. Network modularization by the GN (left) and Infomap (right) algorithms are shown for the crystal structure of glutamyl-tRNA synthetase complexed with tRNA(Glu) and glutamol-AMP (PDB 1N78).

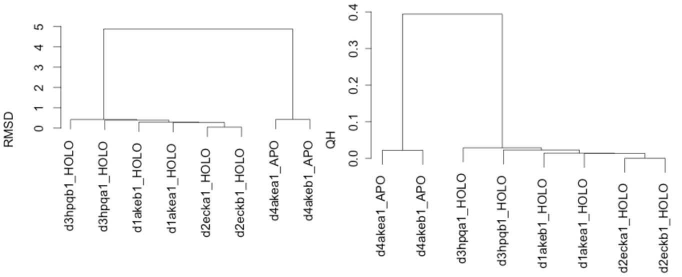


Supp. Fig. 21: Distributions for average pairwise RMSD values across domains within all multiple structure alignments at varying levels of sequence identity.

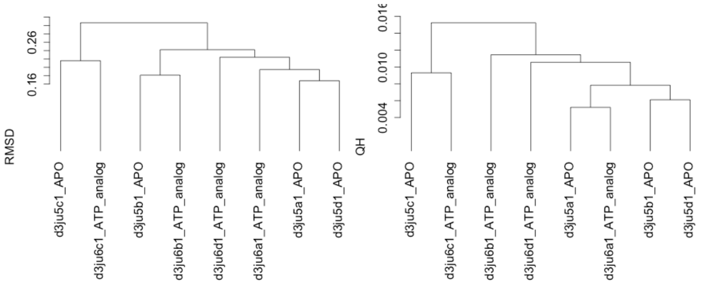


Supp Fig. 22: Intuition behind the k-means algorithm with the gap statistic
The objective is to identify the ideal number of clusters to describe the observed data of 60 points (in blue). This entails defining how well-clustered our observed data appears (given an assigned number of clusters, K) relative to a null model consisting of a randomly distributed set of 60 points (grey) that fall within the same variable ranges as the observed data. Further details are provided by Tibshirani et al, 2001.

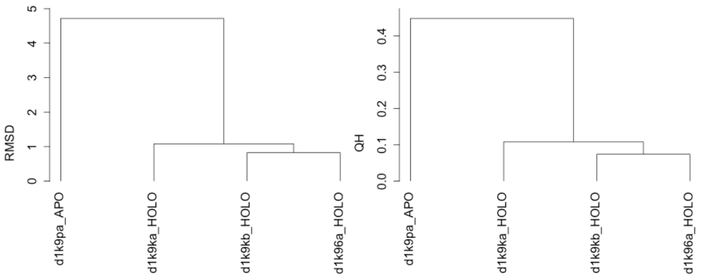
Supp. Fig. 23: Clustering domains based on RMSD generally matches that used when clustering based on Q_H.



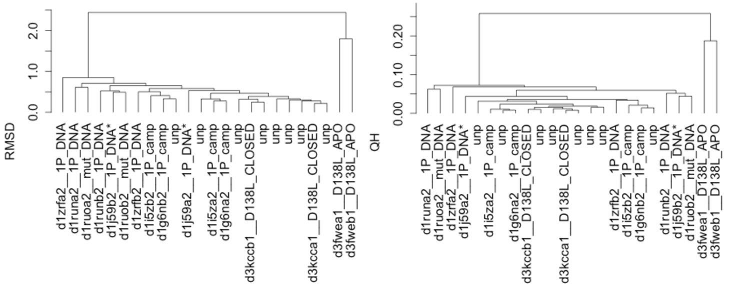
23a) Adenylate kinase



23b) Arginine kinase

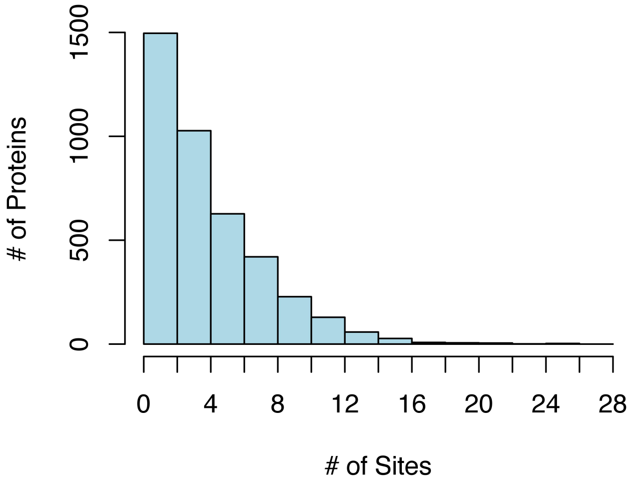


23c) Calcyclin

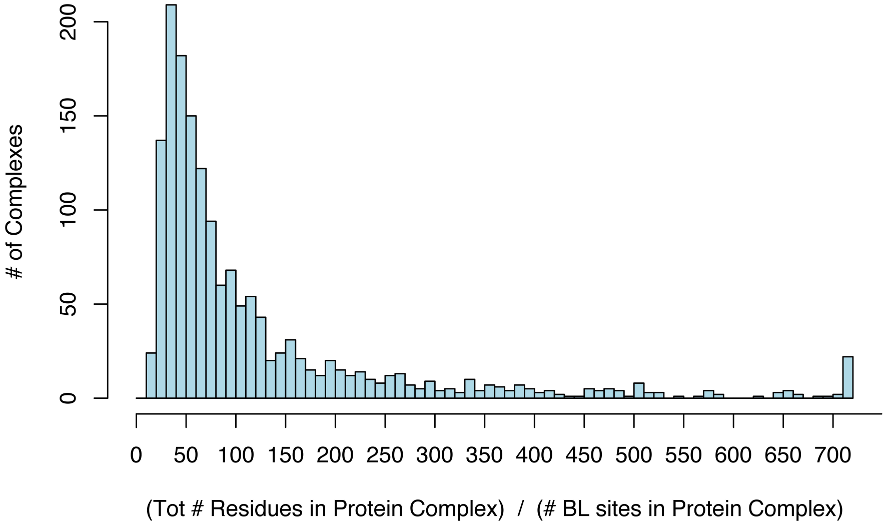


23d) Catabolite activator protein (CAP)

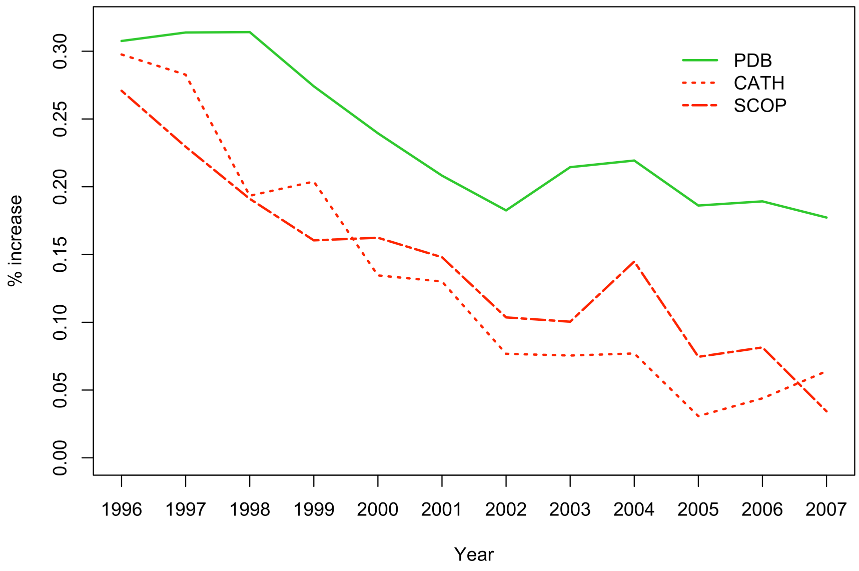
Supp Fig. 24



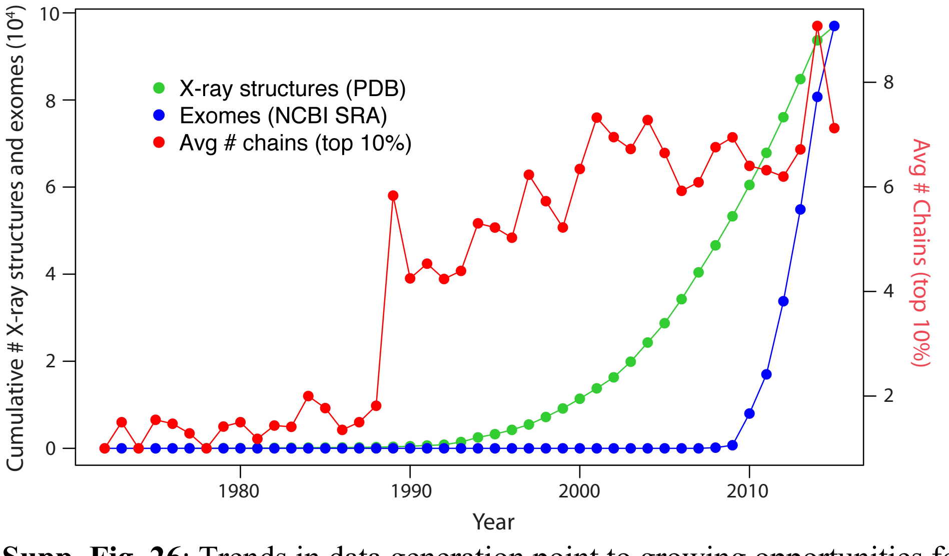
24a) Number of surface-critical sites per protein (PDB chain)



24b) Density of surface-critical sites with respect to number of residues in complex



Supp. Fig. 25: The growth rate of deposited PDB structures from 1996 to 2007, and the concomitant growth rate in the number of folds (as defined by CATH and SCOP). The growing appreciation for dynamic behavior and the importance of conformational heterogeneity is being facilitated by a growing redundancy within the PDB. Such redundancy is represented, for instance, when the same protein is structurally resolved under different conditions, potentially resulting in alternative conformations.



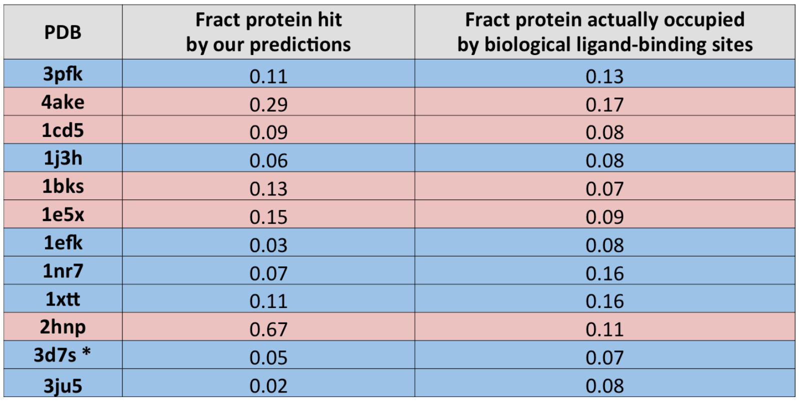
Supp. Fig. 26: Trends in data generation point to growing opportunities for leveraging sequence variants to study structure (and vice versa): The volume of sequenced exomes is outpacing that of structures, while solved structures have become more complex in nature. Red: Average number of chains per PDB (considering the biological assembly PDB files for the top 10% of PDBs for a each year, as ordered by the number of chains for each structure). Green: Cumulative number of X-Ray structures deposited in the PDB. Blue: Cumulative number of exomes stored in the NCBI Sequence Read Archive (SRA). All data were downloaded in May 2015.

HOLO	**APO**
1ake (**AP5**)	4ake
3cep (**G3P, IDM, PLP**)	1bks (**PLP**)
1hor (**AGP**, *PO4*, [& **16G** in pdb 1HOT])	1cd5
2c2b (**SAM**, [& **LLP** in pdb 2c2g])	1e5x
1gz3 (**ATP, FUM**, *OXL*)	1efk (MAK)
1atp (**ATP**)	1j3h
1hwz (**GLU**, **GTP, NDP** [& **ADP** in PDB 1NQT])	1nr7
1xtu (**CTP, U5P**)	1xtt (*ACY*, **U5P**)
1aax **(BPM** [& **892** in PDB 1T49])	2hnp
7at1 (**ATP**, **MAL**, **PCT** [& **CTP** in PDB 1RAC], [& **PAL** in PDB 1D09])	3d7s
3ju6 (**ANP**, **ARG**)	3ju5
6pfk (PGA [& **F6P + ADP** in PDB 4PFK])	3pfk (*PO4*)

Supp. Table 1: Set of 12 canonical proteins, organized by state (*apo* or *holo*). Ligands are given in parentheses (those in bold text designate the ligand used to define residues involved in canonical ligand-binding interactions).

PDB	Fract Protein Hit	Fract Known Bio Sites Hit
3pfk	0.108	1
4ake	0.287	1
1cd5	0.09	0.5
1j3h	0.06	1
1bks *	0.132	0.25
1e5x	0.151	0.667
1efk	0.032	0
1nr7	0.071	0.75
1xtt	0.111	1
2hnp	0.672	1
3d7s *	0.052	0
3ju5	0.017	0
Avgs	**0.15 (0.16 w/o 3d7s)**	**0.60 (0.65 w/o 3d7s)**

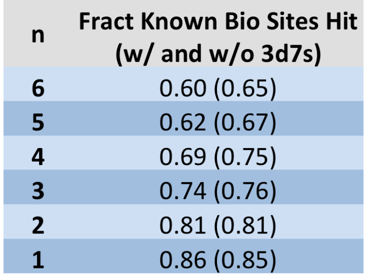
Supp. Table 2: Identifying known ligand-binding sites
The 2nd column designates the fraction of residues that constitute surface-critical residues, and the 3rd column represents, for each structure, the fraction of known ligand-binding sites that strongly overlap with surface-critical sites.



PDB	Fract protein hit by our predictions	Fract protein actually occupied by biological ligand-binding sites
3pfk	0.11	0.13
4ake	0.29	0.17
1cd5	0.09	0.08
1j3h	0.06	0.08
1bks	0.13	0.07
1e5x	0.15	0.09
1efk	0.03	0.08
1nr7	0.07	0.16
1xtt	0.11	0.16
2hnp	0.67	0.11
3d7s *	0.05	0.07
3ju5	0.02	0.08

Supp. Table 3: Do surface-critical sites occupy an exceedingly large fraction of the protein?

For most proteins in the canonical set, the fraction of the protein occupied by surface-critical residues roughly matches the fraction of residues known to be directly involved in ligand binding. For most proteins (blue), the fraction of critical-surface residue is actually lower than that of known ligand-binding residues.



n	Fract Known Bio Sites Hit (w/ and w/o 3d7s)
6	0.60 (0.65)
5	0.62 (0.67)
4	0.69 (0.75)
3	0.74 (0.76)
2	0.81 (0.81)
1	0.86 (0.85)

Supp. Table 4

Here, *n* designates the number of residues within a surface-critical site that overlap with known ligand-binding residues. For the calculations reported above and in the main text, this value is taken to be n=6 (because each surface-critical site typically has 10 residues, and never has more than 10 residues, this criterion enforces that a majority of surface-critical residues within a given site overlap with known ligand-binding residues in order to be counted as a site match). However, as this threshold is relaxed to lower n, the fraction of captured known ligand-binding sites improves rapidly, suggesting that surface-critical sites generally lie close to known ligand binding sites in many cases.

Fraction of Rare SNPs in Critical and non-critical Residues using ExAC

Values outside of parentheses designate results using a rarity threshold of 0.005.
Values within parentheses designate results using a rarity threshold of 0.001.

	Surface-critical	Interior-critical
% of structures such that the fraction of rare SNPs in critical residues is ***greater than*** the fraction of rare SNPs in non-critical residues	9.5 (30.0)	15.5 (41.1)
% of structures such that the fraction of rare SNPs in critical residues is ***LESS than*** the fraction of rare SNPs in non-critical residues	6.0 (13.0)	0 (3.3)

Supp. Table 5

Degree of Concordance Between Community Detection Methods: GN vs. Infomap

Protein (PDB, # residues)	Community Detection Method: GN | InfoMap			
	Modularity	# Comm.	# Critical Residues	% of GN critical residues which match those in Infomap (expected)
tRNA synthetase (1N78, 542)	0.71 | 0.68	14 | 25	47 | 109	0.28 (0.20)
Adenylate kinase (4AKE, 428)	0.73 | 0.70	11 | 20	39 | 82	0.90 (0.19)
Arginine Kinase (3JU5, 728)	0.72 | 0.69	12 | 28	41 | 142	0.22 (0.19)
Tyrosine Phosphatase (2HNP, 278)	0.59 | 0.59	7 | 15	27 | 70	0.26 (0.25)
Phosphoribosyltransferase (1XTT, 846)	0.72 | 0.68	9 | 32	36 | 174	0.22 (0.21)
cAMP-dep. PK (1J3H, 332)	0.66 | 0.64	11 | 19	36 | 78	0.33 (0.23)
Anthranilate synthase (1I7Q, 1418)	0.75 | 0.69	12 | 46	51 | 288	0.31 (0.20)
Malic enzyme (1EFK, 2212)	0.81 | 0.72	17 | 70	74 | 425	0.18 (0.19)
Threonine synthase (1E5X, 884)	0.73 | 0.69	13 | 36	43 | 192	0.28 (0.22)
G-6-P Deaminase (1CD5, 1596)	0.79 | 0.72	18 | 54	58 | 266	0.16 (0.17)
Phosphofructokinase (3PFK, 1276)	0.76 | 0.68	10 | 51	45 | 307	0.24 (0.24)
Tryptophan synthase (1BKS, 1294)	0.77 | 0.69	10 | 46	41 | 284	0.24 (0.22)
Means	***0.73 | 0.68***	***12.0 | 36.8***	***44.8 | 201.4***	***0.3***

Supp. Table 6