doi:10.1016/j.jmb.2004.10.055 J. Mol. Biol. (2004) 344, 1135–1146

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Network Analysis of Protein Structures Identifies Functional Residues

Gil Amitai, Arye Shemesh, Einat Sitbon, Maxim Shklar, Dvir Netanely Ilya Venger and Shmuel Pietrokovski*

Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel

Identifying active site residues strictly from protein three-dimensional structure is a difficult task, especially for proteins that have few or no homologues. We transformed protein structures into residue interaction graphs (RIGs), where amino acid residues are graph nodes and their interactions with each other are the graph edges. We found that active site, ligand-binding and evolutionary conserved residues, typically have high closeness values. Residues with high closeness values interact directly or by a few intermediates with all other residues of the protein. Combining closeness and surface accessibility identified active site residues in 70% of 178 representative structures. Detailed structural analysis of specific enzymes also located other types of functional residues. These include the substrate binding sites of acetylcholinesterases and subtilisin, and the regions whose structural changes activate MAP kinase and glycogen phosphorylase. Our approach uses single protein structures, and does not rely on sequence conservation, comparison to other similar structures or any prior knowledge. Residue closeness is distinct from various sequence and structure measures and can thus complement them in identifying key protein residues. Closeness integrates the effect of the entire protein on single residues. Such natural structural design may be evolutionary maintained to preserve interaction redundancy and contribute to optimal setting of functional sites.

Keywords: protein active sites identification; protein structure analysis; network analysis; closeness degree; ORFans

*Corresponding author

Introduction

Complex systems can be analyzed as networks of interactions between the system components. Analyzing the network can then characterize the whole system and its individual components.[1,2] Protein structures are typically perceived as local structure elements that form diverse topologies and folds.[3,4] However, protein structures can also be represented as networks (graphs) where amino acid residues are the nodes and their interactions are the edges.[5] This approach was used to study various protein aspects, including protein structure flexibility,[6] folding of protein domains,[7] recurring structural patterns,[8] key residues in folding,[9] residue fluctuation,[10] and side-chain clusters.[11]

Identifying functional residues (e.g., active site residues) in proteins is a complex issue, even when atomic detailed structures are available.[12] This is further complicated when no recognizable homologues with characterized functional residues are known. Only a very small fraction of all known proteins has been biochemically well studied. Therefore, for most proteins with solved 3D structures we need *de novo* methods to predict functional residues. Evolutionary conservation together with structure information is successful in predicting some ligand binding and active sites in various proteins.[13–16] However, some proteins with known structures do not have determined, and maybe even existing, homologues. Using only structural data, functional residues were identified by computing structure energetics and ionization properties.[17,18] In a different approach, such residues could be characterized by their structural

Abbreviations used: RIGs, residue interactions graph(s); RSA, relative solvent accessibility; MCC, Matthews correlation coefficient.

E-mail address of the corresponding author: shmuel.pietrokovski@weizmann.ac.il

properties or recurring structural motifs.[8,19] Combining different structural and evolutionary residue properties improves the identification of active site residues.[20] Each method has its own advantages and limitations, but even integrating different approaches could not identify all sought-after sites. New ways to characterize and predict key protein residues are still needed.

The interactions of protein residues within and between functional sites are crucial for protein activity. Analysis of protein structures found them to be small-world networks.[9,10,21] Such networks are characterized by both clusters of local interactions and "long range" interactions between different clusters.[22] Frequently these networks include a small number of central nodes that are hubs through which many nodes can indirectly connect.[23] We sought to examine whether central nodes of protein residue interaction networks correspond to functional residues. The closeness of a residue to other residues on the network was found to characterize many functional protein sites. We preformed a large-scale prediction of active site residues, and examined the relation of residue closeness to other residue functions. Active site residue prediction relied on single structure analysis without using homologous sequences or structures. Residue closeness is distinct from other structurally derived residue properties. However, high closeness is associated with sequence conservation and is characteristic of key positions in general, where mutations can abolish protein activity. It can thus be used, with other measures, to analyze protein structures. This includes structures with unique sequences or folds. We discuss here possible explanations and consequences for the relation between residue functionality and high network closeness.

Results

Transforming protein structures into mathematical graphs

Protein structures were transformed into mathematical graphs (networks) by identifying all interactions between the amino acids in each structure. We first found all interatomic contacts using the CSU program[24] and then integrated these contacts by amino acids. In the resulting residue interactions graphs (RIGs) amino acid residues are the nodes and their interactions are the edges. We set out to characterize properties of individual amino acid residues (nodes) based on their relation with other residues of the protein (network). For this end we examined the centrality of network nodes. Network centrality measures of individual nodes are based on their distance from other nodes, or on their position along paths between other nodes.[25] The distance between two nodes is the number of edges along the shortest path between the nodes. Centrality measures can also be classified by their dependence on the whole or on the local structure of the network. Degree-centrality is the number of edges each node has (i.e., its number of immediate neighbors), and is thus a local measure. Closeness-centrality is derived from the mean distance of a node to all other nodes, and is thus a global measure. We first compared these two distinct centrality measures of protein RIGs.

Degree and closeness centrality measures were examined for all amino acids in a representative group of 178 protein chains of enzymes with known structure and uniformly defined active sites.[19] There is a positive correlation between the two measures for the entire set of amino acid residues ($r = 0.59$). Amino acid residues with low degree and closeness values are mainly highly exposed (relative solvent accessibility [RSA] $\geq 50\%$), while unexposed amino acids (core, RSA $= 0$) have mainly high

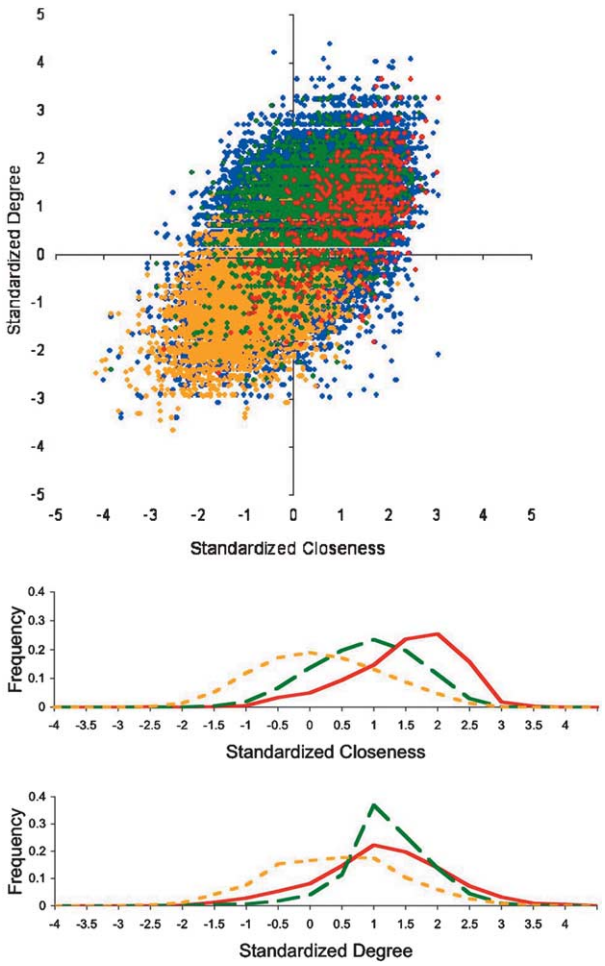

Figure 1. Enzyme residues closeness and degree. Values are shown for all 59,935 residues from the 178 enzyme chains. On top, highly exposed residues (RSA > 50) are in orange, core residues (RSA = 0) are in green, active site residues are in red, and all other residues are in blue. In the distributions below, exposed residues (RSA > 0) are in orange dotted lines, core residues are in green broken lines, and active site residues are in red. Standardized closeness and degree values are calculated as noted in Methods section.

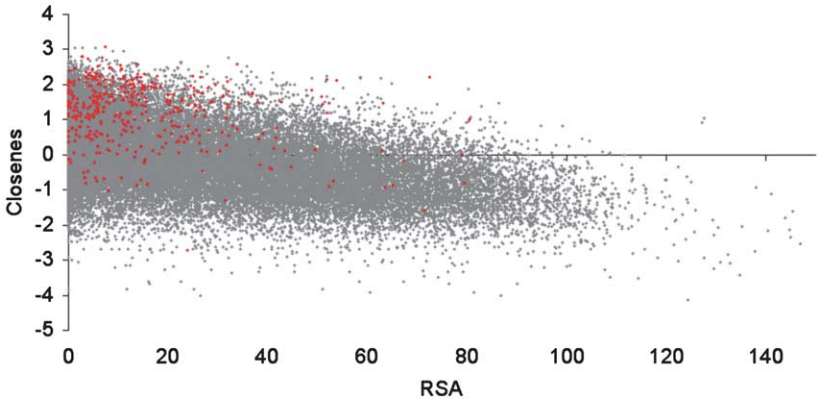

Figure 2. Closeness and accessibility of enzyme active sites. In red are the 567 active site residues and in grey all other residues from the 178 enzyme chains.

values for both centrality measures. Most active site residues have high degree and closeness values (Figure 1). Closeness values of active site residues are significantly higher than those of the other residues, including those in the protein core. Degree values of active sites are less distinct than closeness values from the corresponding values of core and surface residues (Figure 1). Degree measure is also similar to other measures that are calculated from the local environment of structure residues (e.g., solvent accessibility, depth). In contrast, closeness is an attribute of single residues that depends on all of the protein residues.

The independence of the closeness measure was examined by comparing it to other residue structural properties. Both closeness and RSA values characterize active sites, but they are only partially related (correlation of $r = -0.32$, Figure 2). In addition, there is no detectible relation between residue closeness values and either properties of protein structure clefts they are found in (size and rank), or their *B*-factors (mean thermal motion of all the residue atoms) (Supplementary information). Finally, ROC curve analyses[26] of enzyme active site predictions using RSA and using RSA and closeness together, show a very significant improvement when adding closeness to RSA (Supplementary information). These findings show that closeness is a distinct and informative property of protein residues. Our results indicated that active sites can be characterized and predicted by their closeness centrality values, with or without other structural properties.

Prediction of active site residues by the closeness parameter

RIG closeness and RSA values were used to predict the active sites in the group of 178 previously characterized enzyme chain structures. In this initial analysis we simply used a closeness lower-limit value and an RSA lower and upper-limit values. Many combinations (>2000) of threshold values were evaluated by a jackknife procedure. This objectively identified a set of optimal parameters to use for active site prediction (standardized closeness values ≥ 1.1 and RSA values of 4.5–40%). These parameters gave over the entire data set an average of 46.5% sensitivity (fraction of identified active site residues) and 9.4% specificity (fraction of correct prediction). The median overall performance rate (MCC, Matthews correlation coefficient, a measure that integrates sensitivity and specificity) of the prediction was 0.20. These values are comparable to the best published active-site predictions from structural data alone.[20] Full correct prediction (100% sensitivity) was found for about quarter of the proteins (42/178) and partial correct prediction was found for 70% of the proteins (Figure 3(A)). These are highly significant results giving a *Z*-score of 185.2, highly unlikely to occur by chance, $P \sim = 0$. Better prediction was found for large proteins, longer than 120 amino acid residues, (49.0% sensitivity and 9.7% specificity means) than for small proteins, ≤ 120 amino acid residues, (13.2% sensitivity, 4.7% specificity). No predictions at all (zero sensitivity and specificity) were found for nine of the 12 small proteins in the examined data. A WWW server implementing this method is available†.

Relation between residue closeness values and functionality

Our results showed that some residues have high closeness values although they are not classified as active site residues according to the definition used by Bartlett *et al.*[19] Nevertheless, some of these residues might be considered as active sites using other criteria, or might have other functional importance (e.g., binding of substrates, of cofactors, or of metals). We chose to examine the relation between closeness values and sequence conservation for representative protein families. Sequence conservation is a good indicator for important sites on proteins. Closeness was found to positively correlate with sequence conservation, with values ranging from 0.24 to 0.62 (Supplementary information).

Predicting enzyme active site residues by integrating closeness and sequence conservation significantly improved prediction by only using

† http://www.weizmann.ac.il/SARIG

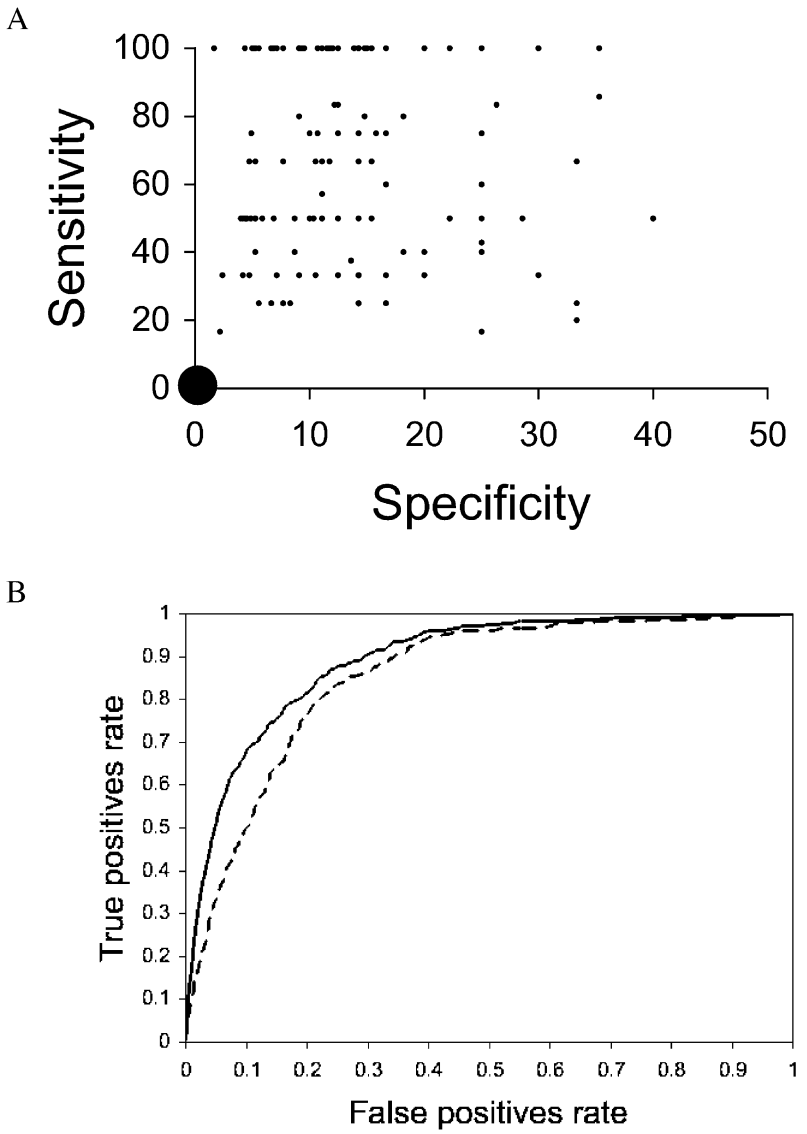

Figure 3. Active site residues prediction using closeness, RSA and sequence conservation. (A) Active site prediction success for all analyzed 178 chains using integrated closeness and RSA. Large filled circle indicates 54 chains that had 0% sensitivity and specificity. (B) ROC curves for all residues from the analyzed chains, using sequence conservation active site prediction (broken line) and using integrated sequence conservation and closeness (continuous line). Sequence conservation values are from the CONSURF server[60] output for each chain. Closeness standardized values and CONSURF sequence conservation values were integrated for each residue by adding the closeness value to the conservation value multiplied by −2.

sequence conservation. For highly specific predictions (5% false positives) using only sequence conservation only gave 33% sensitivity. Adding closeness to sequence conservation increased the sensitivity to 51% for the same specificity value (Figure 3(B)).

Large-scale experimental data on the relation between protein activity and amino acid substitutions are available for a few proteins. Typically most of the protein residues were mutated one at a time to various other amino acids, and the effect on protein activity measured.[27–32] Such exhaustive mutagenesis studies examined both conserved and variable residues. The positions where mutations effected protein activity were identified directly in T4-lysozyme and barnase,[29,32] and indirectly in TEM1-β-lactamase (TEM1).[27] Altogether, the influence of each position on protein activity was roughly estimated due to several reasons. First, the activity threshold for enzyme inactivation is different in the three examples (0.1, 3, 27% for barnase, T4-Lysozyme, and TEM1-β lactamase, respectively). Second, not all possible amino acid substitutions were examined in these studies. We thus considered positions where mutations could abolish activity as key positions.

Key positions in all three structures had higher RIG closeness values than positions where mutations had no apparent effect on protein activity (Figure 4). To quantitatively estimate the use of closeness and compare it with the use of RSA to identify key positions we performed a logistic regression analysis. We combined the data for the three structures, but only examined the surface exposed residues, avoiding the simple task of identifying key residues in protein cores. The log-odds estimate was 3.7 for using closeness, and 0.9 for using RSA. Thus, for exposed residues closeness has a strong predictive power while RSA, a typically used measure,[27–30] is a weak predictor (log-odds estimate close to one).

We correctly predicted both evolutionary

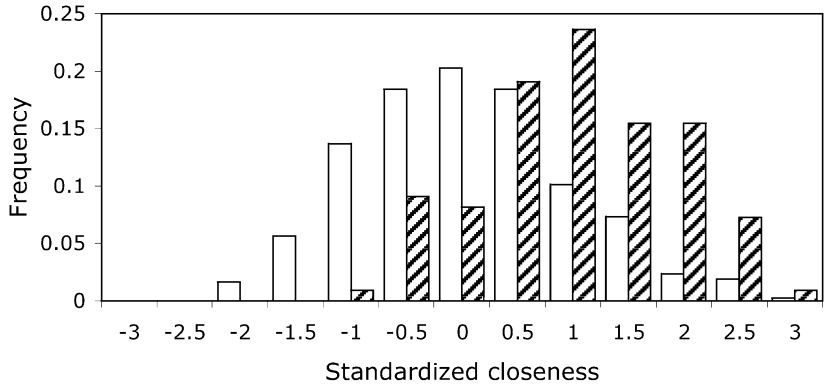

Figure 4. Closeness distribution in key and mutation-tolerant residues. Closeness values of T4-lysozyme, barnase, and TEM1-beta-lactamase (PDB accessions 1lzm, 1a2p, 1axb, respectively). Values of mutation-tolerant residues are shown in empty bars and hatched bars show the values of key residues, residues where mutations could abolish protein activity.

conserved- and variant key positions in TEM1. Seventy percent of the key positions in this enzyme are variant (30/43).[27] More than a third of these key residues have high closeness values (11/30). This fraction of variant key residues that have high closeness is significantly higher than expected ($p = 5\times10^{-3}$) from the total number of residues with high closeness in the protein (17%, 46/263). Similar significance ($p = 6\times10^{-3}$) is found even if we only consider the TEM1 solvent exposed residues. We also identified T4-lysozyme variable key residues by their closeness values (not shown). This could not be tested in barnase since all its key residues are conserved (not shown). Closeness values can thus complement sequence conservation in identifying key positions.

RIG closeness of functionally important non-catalytic residues

To complement the large scale and general analyses on the nature of residues with high closeness values we conducted more thorough examination of specific protein structures. We chose well-studied examples to study substrate binding sites, allosteric sites and the effect of regulatory post-translational modifications.

Substrate, cofactor and ion binding sites

In subtilisin protease,[33] high closeness values identified both substrate binding and one of the two cation binding-sites, in addition to all catalytic site residues. While the substrate and catalytic sites are close to one another, the identified cation binding-site is on the opposite side of the protein structure (Figure 5). The cation binding site stabilizes subtilisin structure.[34] We also identified the substrate and ion binding sites, together with active site, in the ERK2 MAP kinase.[35] ERK2 ATP and Mg^{2+} binding sites could be found by their high closeness values (Figure 6(A)). Acetylcholinesterase (AChE) catalytic site is situated within a deep narrow gorge.[36–39] All residues within the gorge including the catalytic triad, oxyanion hole and the anionic site had high closeness values. AChE

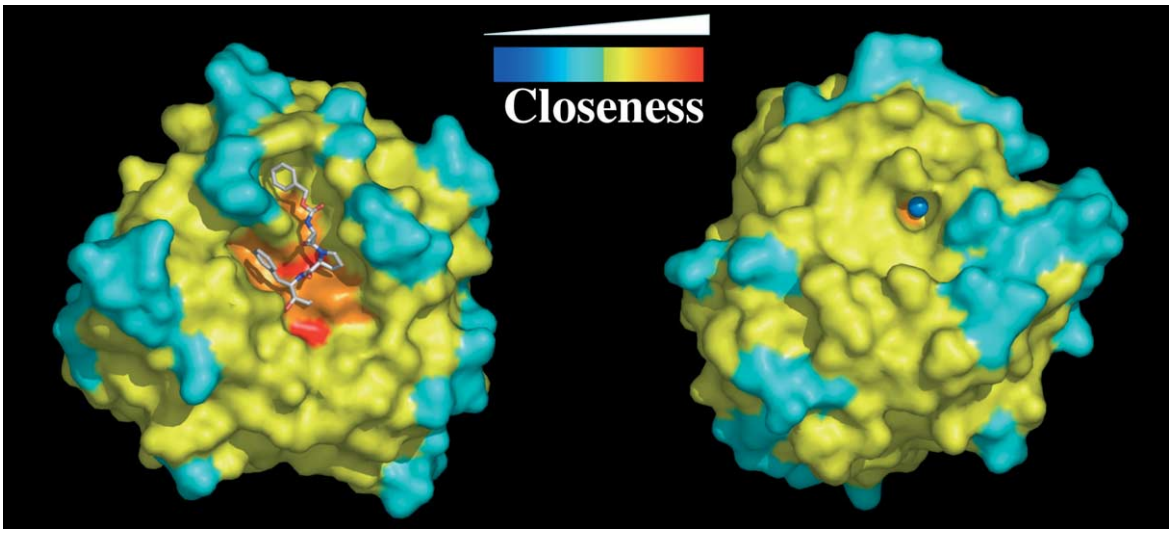

Figure 5. Closeness analysis of subtilisin DY protease. Closeness residue values are shown on the surface of the protein (PDB accession 1BH6). Closeness increases from blue to red. The left view shows the protease active site with a synthetic inhibitor, shown in sticks. The right view is related to the top by about 90° counterclockwise turn on the *Y* axis. It shows a Na atom in cation binding site B. Note the infrequency of residues with high closeness values and their exact overlap with the subtilisin active and cation binding sites.

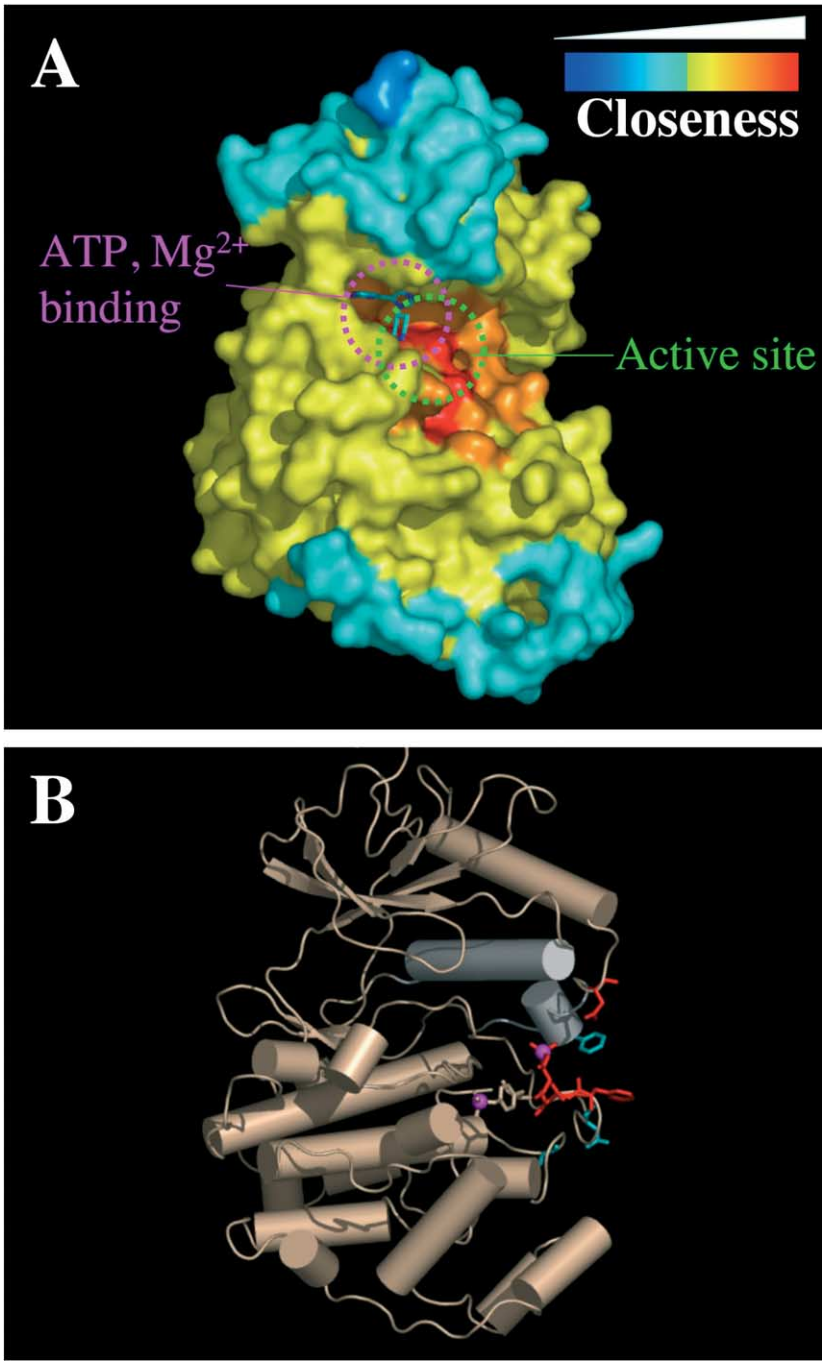

Figure 6. Closeness values of ERK2 MAP kinase. (A) Surface representation of unphosphorylated ERK2 MAP kinase (3ERK) bound to an inhibitor in the ATP binding site (shown in sticks). Residues are colored by closeness values, with red and blue colors corresponding to the highest and lowest closeness values, respectively. The active site and ATP-Mg2+ binding region have high closeness values. (B) Closeness value changes between phosphorylated (P-ERK2) and unphosphorylated ERK2 (ERK2) forms shown on the structure of P-ERK2 (2ERK). Residues whose closeness significantly changed (more than one standard deviation) in the P-ERK2 compared with ERK2 are shown as sticks. Significantly increased closeness is shown in red and significant decrease is shown in cyan. Helix-C is represented by the large grey cylinder, and the helix that forms upon phosphorylation in the L16 region is represented by the small grey cylinder. Phosphate groups are shown in magenta. The structure is oriented as in (A).

peripheral anionic site residues, located on the rim of the catalytic site gorge, and acyl-binding pocket also had high closeness values (Supplementary information).

Allosteric sites

Glycogen phosphorylase (GP) was the first allosteric enzyme to be isolated and analyzed in detail.[40] AMP binding or phosphorylation mediate the transition between GP active and inactive structural states. Long-range allosteric changes occur between the GP catalytic site and either AMP-binding or phosphorylation site (45 Å and 35 Å apart, respectively). The transition into the activated state is initiated by sliding of two helices, found in the interface of the GP homodimer. This modifies a β-sheet in each of the subunits, that in turn transmits the structural change to the catalytic site.[41]

A RIG analysis of activated dimers shows that the residues of the AMP-binding allosteric site have high closeness values (Mean closeness 1.2 ± 0.16). Residues of a different, recently discovered, allosteric site[42] have even higher closeness values (mean standardized closeness 1.7 ± 0.23). Moreover, 20 out of 26 residues mediating the conformational changes[43] have high closeness values as well (mean standardized closeness of all 26 residues $1.39 \pm$

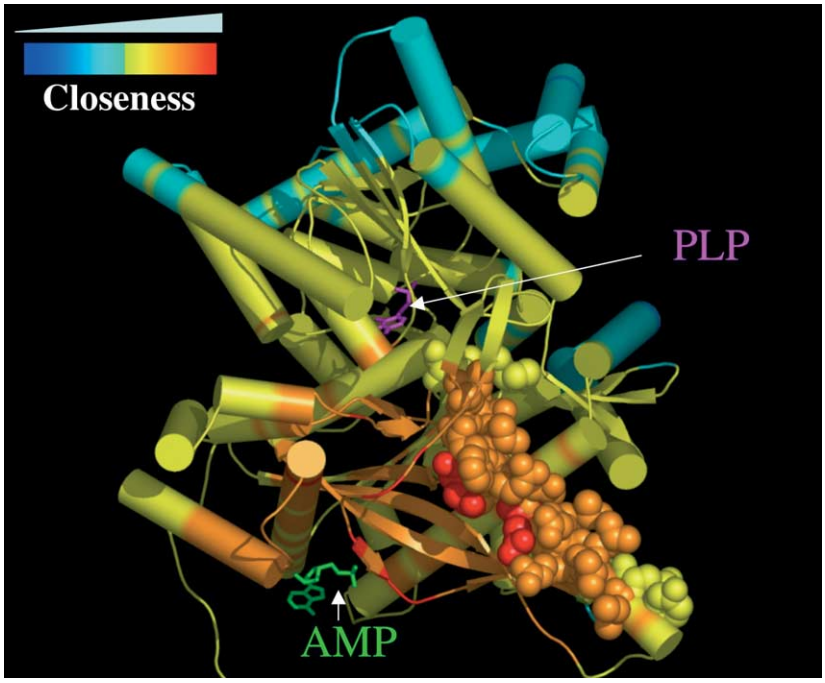

Figure 7. Closeness and allosteric transition glycogen phosphorylase (GP). Chain A of activated (R-state) GP functional dimer[43] (PDB accession 7GPB) is shown with residues colored by closeness values. Residues that transmit the allosteric signal occurring upon AMP binding are shown as spheres (residues 133, 162–165, 262–279, 281). Pyridoxal phosphate cofactor (PLP), situated within the catalytic site, is shown in magenta. AMP bound to its regulatory site is shown in green. Closeness values were calculated by transforming GP dimers into a single network.

0.45). This is highly significant relative to the number of high closeness residues in the whole protein (159/824, $p=2\times10^{-9}$). Moreover, all residues of the AMP binding site form a continuous stretch of high closeness amino acid residues that reach the residues that structurally change upon AMP binding (Figure 7).

MAP kinase activation and RIG closeness changes

MAP kinase isoforms ERK2 and ERK1 are well studied ubiquitous, growth-factor-activated, tyrosine-phosphorylated, enzymes.[44–46] ERK2 reaches maximal activity upon phosphorylation of two residues (T183 and Y185).[46–48] We compared residue closeness value changes between the inactive ERK2 unphosphorylated[35] and active phosphorylated (ERK2-P2)[49] forms. Phosphorylation causes local and global structural changes in many ERK2 regions, including rotation of a whole domain.[49] However, large increase in closeness (>1 closeness standard deviations) is found in a few residues forming the activation-loop (F181, L182, T183), and residue D335. The activation-loop contains the two phosphorylation sites. Upon phosphorylation the loop refolds and activates the protein. The loop contacts the L16 region, which includes D335. The conformational changes in L16 are coordinated with position shifts helix C into the active site[49] (Figure 6). The largest increase in closeness (1.76 standard deviations) is in the phosphorylated residue T183. This residue is the hub of two new hydrogen and ionic interaction networks formed within ERK2-P2. It is responsible for bringing the active site residues into alignment, the ATP binding Helix C closer to the active site, and improving the interaction between the conserved catalytic residues K52 and E69.[49] The largest decrease in closeness values (−1.1 standard deviations) is identified in F329. It is in a part of L16 that forms a helix after phosphorylation, and interacts with the activation loop.

Discussion

Transforming protein structures into networks of residue interactions allowed us to examine the relationship between each amino acid residue and the protein structure as a whole. Relationships were evaluated by the centrality of each residue within the interaction network. In this study we used the closeness parameter as a measure of network centrality, where highly central residues have high closeness values. Such residues interact with most other residues directly or by a few intermediates. Residues with high centrality values are then assumed to efficiently integrate and transmit information from and to the rest of the protein. Such information may appear in a chemical or physical form, affecting the attraction or repulsion of amino acids. Our results demonstrate that many active site residues have high centrality values (Figure 2). The high centrality of active site residues suggests they can effectively (directly) disseminate and receive signals from the rest of the protein.

Active site residues are highly central relative to solvent exposed residues, where active sites are typically found, and also even relative to the core of the protein, where one would expect residues to be highly central (Figure 1). Active site residues can be identified by various other approaches. Some approaches rely on sequence conservation deduced from multiple sequence alignments,[50] some on structural features such as electrostatic[17,18] or structural clefts[51] and some on a combination of

such features.[13,19,20,52] Here we report that the combination of centrality and solvent accessibility already enabled us to identify active site residues in 70% of the proteins within a data set of 178 representative enzymes generated by Bartlett *et al*.[19] (Figure 3(A)). This is a large scale analysis, comprising approximately 60,000 residues and 500 active site residues. Our prediction results are thus highly significant.

Our method defines individual active site residues, not just patches including them. In addition, other residues, not directly linked to catalysis (e.g., binding cofactors or substrates), can also be identified using our approach (Figures 5 and 6 and Supplementary information). We suggest that our new approach for active site identification not only complements other known methods but also leads to a deeper understanding of the relationship between the whole structure of proteins and their specific functional sites. This relationship is probably a direct consequence of the necessary functional robustness of proteins.

Robustness to environmental and mutational perturbations is fundamental to protein function. Proteins hence evolve toward a common design that supports such features. We found this design reflected in the centrality of amino acid residues within residue interaction networks. One result of such network design is that many residues contribute to the optimal arrangement of the active site (e.g., orientation, charge). This can yield protein activity resilient to environmental perturbations (e.g., changes in pH, temperature, ionic strength, and mutations) by preserving the essential state of the active site. In each protein, many network edges (interactions) or nodes (residues) can be individually removed while the remaining nodes will continue to interact by alternate paths. Such a model is supported by data from exhaustive mutagenesis studies, where numerous single-site substitutions are tested within each position of the protein. These studies show that most mutations do not impair protein activity.[27,29,30–32] We found that mutations in highly central residues often impair activity (Figure 4). These results are remarkable since we could only use crude estimates for the effect of the mutation on the activity of each protein.

Typically, only a few amino acid residues within the protein are directly involved in catalysis.[19] Nonetheless, protein activity can be modulated by non-catalytic residues, including residues very distant from the active site. This is observed in allosteric enzymes where effectors often bind to residues distant from active sites. Our analysis of glycogen phosphorylase, a well-studied allosteric enzyme, shows that both allosteric regulatory sites and the residues involved in the allosteric-derived structural changes are central within the residue network of this enzyme. This supports our suggestion that the interactions between allosteric and active sites lead to the high centrality of both.

Amino acid sequence conservation is often attributed to important sites such as those directly related to protein function.[13,50,52] We followed our findings on the relation between network centrality and functional residues by evaluating the relation between network centrality and sequence conservation. Our results show that centrality and sequence conservation are positively correlated in several diverse protein families (Supplementary information). Sequence conservation identifies residues that are under the same selection pressure in the whole family. However, some residues that are particularly adjusted for the function or structure in specific proteins within the family will not be conserved. This is shown in TEM1-β-lactamase where an exhaustive mutagenesis study showed that some positions are both completely intolerant to mutations and non conserved within their protein family.[27] Using network centrality we identified such residues, together with conserved residues. Thus, for proteins with known structure, our approach may complement sequence information in the prediction of non conserved functional residues.

Active sites were correctly predicted for the majority of the enzymes in our large-scale analysis (70% of the structures, that had a median of 67% sensitivity and 13% selectivity). No correct predictions at all, however, were found by our method for the remaining enzymes. Examining this group showed that it included nine of the 12 smallest enzymes ($\leq$120 residues). Residue interaction networks of such proteins may not be amenable to our analysis approach due to their small size. Alternatively, their active site design may be different from that of larger enzymes. These sites might be identified by analyzing their networks in a different fashion. We are yet uncertain what might be common for the remaining (large) enzymes with no correct predictions. One distinguishing feature is the distribution of catalytic function types.[19] This distribution is significantly different between the predicted and unpredicted active site residues in large enzymes ($p = 2 \times 10^{-2}$). Two types of catalytic functions (primer and acid–base, as defined by Bartlett *et al*.[19]) contributed most of the difference, indicating that our approach and current thresholds are more successful in predicting certain types of active site residues.

Improvement in active site prediction was achieved by combining all chains of some multimeric enzymes into a single interaction network. However, this was not noted for all such enzymes that were examined, and the overall prediction success did not noticeably improve, relative to single chain network analysis (not shown). We believe that prediction success depends on the location of the active site residues. Using all chains gave highly successful predictions for active sites with residues in more than one subunit, but not when all the active site residues were within a single subunit. It is therefore advantageous to include all available biochemical knowledge

upon deciding on the analysis parameters and interpreting its results.

In summary, our results show that key protein positions have high centrality values in residue interactions networks. Catalytic, substrate and co-factor binding, and mutation intolerant residues were found highly central. Other types of functionally and structurally important residues may yet be found to have specific interaction networks features. Our approach is unique in not focusing on residue physicochemical properties or on their evolutionary conservation. Rather, it relies on the interactions of each residue and the rest of the protein. This type of analysis can complement other existing methods for studying proteins (Figure 3(B) and Supplementary information). Our active site residue predictions, using only closeness and RSA values (47% sensitivity, 9% specificity), are comparable to the structural based findings of a recent work that analyzed the same enzyme set, integrating RSA, depth, secondary structure, cleft features, and amino acid type (41% sensitivity, 10% specificity).[20] Using interaction networks may be particularly useful for protein structures having no known homologous sequences (ORFans).[53,54] We also showed that non-conserved key positions can be identified by network analysis. This is important since while sequence conservation often indicates crucial protein positions, such positions can also be variable.[55] Protein network analyses can thus form the basis for identifying important species-specific and allele specific residues. Not depending on prior training or comparison to known protein sequences and structures, our approach is a *de novo* prediction method. As such, it may also uncover functional sites dissimilar to known sites.

Protein structures provide detailed three-dimensional information of the proteins. However, it is not obvious how and which of these data identifies the function of each protein and its specific functional sites. A reductionist approach, transforming protein structures into abstract networks, provides a distinct depiction of proteins. We found that a basic centrality measure of networks nodes can predict functionally important protein residues. Our approach enhances other methods of protein function analysis and illustrates key aspects in natural design and evolution of proteins.

Methods

Transforming protein structures into residue interaction graphs

For each protein chain all interatomic contacts were found using the CSU program.[24] These atomic contacts were integrated for each amino acid residue. Residue interactions form the edges, and their connected residues form the nodes of the protein RIG. Interactions included the backbone peptide bonds as well as non-covalent bonds, such as hydrogen and hydrophobic interactions.

Protein structures analysis

Protein structures were obtained from the PDB database†.[56] The structure set used for the large scale examination of our approach is that described by Thornton and co-workers.[19] It does not contain similar (homologous) pairs and covers all six top-level enzyme classification (EC) numbers.[57] Residue relative accessibility was calculated by the NACCESS program.[58] Protein structures are shown with the PyMol program (http://www.pymol.org).

Multiple sequence alignment

Structure and sequence alignments of protein sequences of similar structures were obtained from the HOMSTRAD database[59] or calculated by us. Conservation of multiple sequence alignment positions was calculated by the CONSURF server.‡[60] The server calculated the conservation for 91% of the 178 analyzed chains, not identifying enough homologs for the following 16 structures: 1c3j, 2thi, 1gog, 2plc, 3csm, 4kbp, 1chk, 1d8h, 1pya, 1coy, 1fui, 1pgs, 1ps1, 1vnc, 2cpo, and 1uox.

Network analysis

Analysis of protein structure residue interactions networks was carried out using programs specifically written in Perl v5.6.1 and Java v1.4. Graph representations and shortest paths of the residue network were calculated using the JDSL java module (Copyright © 1999, 2000 Brown University, Providence, RI and Algomagic Technologies, Inc., Belmont, MA). Network centrality measures were developed by Freeman, Beauchamp, and Sabidussi.[61–63] The formula for closeness centrality calculation[62,63] is shown in equation (1).

Basically "closeness centrality" of node $x(C(x))$ is calculated as follows:

$$C(x) = (n-1)/\sum_{y \in U, y \neq x} d(x,y) \quad (1)$$

Where $d(x,y)$ is the geodesic distance (shortest-path) between node x and any node y. U is the set of all nodes and n is the number of nodes in the network. The closeness value is therefore the inverse of the average distance between x and other nodes ($\bar{d}$) (equation (2)).

$$C(x) = 1/\bar{d} \quad (2)$$

Closeness values are then standardized by calculating their standard deviation from the mean value in each protein structure.

Measuring cleft volume and rank

Protein structure cleft volumes were calculated by the CASTp server§[64] using 42 representative proteins from different enzyme class, subclass, and sub-subclasses (first three EC number fields).

Statistic procedures

Logistic regression analysis was done using the SAS software version 8.02 (SAS institute Inc., Cary NC, USA), using default parameters.

† http://www.rcsb.org/pdb/
‡ http://consurf.tau.ac.il
§ http://cast.engr.uic.edu/cast/

Probabilities for finding residues with high closeness values (>1) were estimated using hypergeometric distribution. Significance of active site prediction was assessed by the method used by Aloy *et al*.[13] considering all 59,935 residues from the 178 analyzed chains. The *Z*-score was thus computed with $n=59935$, $P_O=0.082$, and $P_R=0.0094$ as

$$z = (P_O - P_R)/(P_R(1 - P_R)/n)^{0.5}$$

Identifying optimal parameters for active site prediction

A Jackknife procedure was used to identify optimal prediction parameters. For each 178 analyzed chains the parameters were chosen from the success rate (MCC) of the other 177 chains. The median MCC of each thresholds combination was calculated for all 177 chains. The combination with the highest median was used to predict the active sites for the current analyzed chain. Two thousand one hundred and ninty seven threshold combinations were examined as follows: standardized closeness lower thresholds between 0.9 and 2.1 in intervals of 0.1, RSA lower thresholds between 2.5 and 8.5 in intervals of 0.5, RSA upper thresholds between 33 and 45 in intervals of one.

Standardizing the degree of connectivity

To standardize the degree of connectivity for each amino acid type we collected data of amino acid connectivity from over 2000 structures (~500,000 amino acid residues). These representative structures were retrieved from the PDB-REPRDB database,[65] and had the following properties: less than 30% sequence identity between the proteins, resolution <3 Å, *B*-factor <0.3, no chain breaks, no mutants and no membrane proteins. The degree of residue connectivity was extracted from each protein using a self-written program combined with the CSU program.[24] Connectivity degree was then standardized as number of standard deviations from the mean.

Acknowledgements

We thank Joel Sussman, Israel Silman, Miri Eisenstein, Debbie Fass, Gideon Schreiber, Sven Bergmann for encouragement and insightful discussions, Edna Schechtman for expert statistical consulting, Vladimir Sobolev for developing and providing the CSU software, and Gail Bartlett for providing active site data.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2004.10.055

References

1. Strogatz, S. H. (2001). Exploring complex networks. *Nature*, **410**, 268–276.
2. Albert, R. & Barabasi, A. L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 50–97.
3. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
4. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
5. Lifson, S. & Sander, C. (1979). Antiparallel and parallel beta-strands differ in amino acid residue preferences. *Nature*, **282**, 109–111.
6. Jacobs, D. J., Rader, A. J., Kuhn, L. A. & Thorpe, M. F. (2001). Protein flexibility predictions using graph theory. *Proteins: Struct. Funct. Genet.* **44**, 150–165.
7. Dokholyan, N., Li, L., Ding, F. & Shakhnovich, E. (2002). Topological determinants of protein folding. *Proc. Natl Acad. Sci. USA*, **99**, 8637–8641.
8. Wangikar, P. P., Tendulkar, A. V., Ramya, S., Mali, D. N. & Sarawagi, S. (2003). Functional sites in protein families uncovered *via* an objective and automated graph theoretic approach. *J. Mol. Biol.* **326**, 955–978.
9. Vendruscolo, M., Dokholyan, N. V., Paci, E. & Karplus, M. (2002). Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **65**, 061910.
10. Atilgan, A. R., Akan, P., Baysal, C., Vendruscolo, M., Dokholyan, N. V., Paci, E. & Karplus, M. (2004). Small-world communication of residues and significance for protein dynamics. Small-world view of the amino acids that play a key role in protein folding. *Biophys. J.* **86**, 85–91.
11. Kannan, N. & Vishveshwara, S. (1999). Identification of side-chain clusters in protein structures by a graph spectral method. *J. Mol. Biol.* **292**, 441–464.
12. Jones, S. & Thornton, J. M. (2004). Searching for functional sites in protein structures. *Curr. Opin. Chem. Biol.* **8**, 3–7.
13. Armon, A., Graur, D. & Ben-Tal, N. (2001). ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **307**, 447–463.
14. Aloy, P., Querol, E., Aviles, F. X. & Sternberg, M. J. (2001). Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* **311**, 395–408.
15. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.
16. Landgraf, R., Xenarios, I. & Eisenberg, D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* **307**, 1487–1502.
17. Ondrechen, M. J., Clifton, J. G. & Ringe, D. (2001). THEMATICS: a simple computational predictor of enzyme function from structure. *Proc. Natl Acad. Sci. USA*, **98**, 12473–12478.
18. Elcock, A. H. (2001). Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.* **312**, 885–896.

19. Bartlett, G. J., Porter, C. T., Borkakoti, N. & Thornton, J. M. (2002). Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **324**, 105–121.
20. Gutteridge, A., Bartlett, G. J. & Thornton, J. M. (2003). Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J. Mol. Biol.* **330**, 719–734.
21. Greene, L. & Higman, V. (2003). Uncovering network systems within protein structures. *J. Mol. Biol.* **334**, 781–791.
22. Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, **393**, 440–442.
23. Barabasi, A. L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**, 509–512.
24. Sobolev, V., Sorokine, A., Prilusky, J., Abola, E. E. & Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**, 327–332.
25. Freeman, L., Borgatti, S. & White, D. (1991). Centrality in valued graphs: a measure of betweenness based on network flow. *Social Netw.* **13**, 141–154.
26. Metz, C. E. (1978). Basic principles of ROC analysis. *Semin. Nucl. Med.* **8**, 283–298.
27. Huang, W., Petrosino, J., Hirsch, M., Shenkin, P. S. & Palzkill, T. (1996). Amino acid sequence determinants of beta-lactamase structure and activity. *J. Mol. Biol.* **258**, 688–703.
28. Markiewicz, P., Kleina, L. G., Cruz, C., Ehret, S. & Miller, J. H. (1994). Genetic studies of the lac repressor. XIV. Analysis of 4000 altered *Escherichia coli* lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *J. Mol. Biol.* **240**, 421–433.
29. Rennell, D., Bouvier, S. E., Hardy, L. W. & Poteete, A. R. (1991). Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.* **222**, 67–88.
30. Terwilliger, T. C., Zabin, H. B., Horvath, M. P., Sandberg, W. S. & Schlunk, P. M. (1994). *In vivo* characterization of mutants of the bacteriophage f1 gene V protein isolated by saturation mutagenesis. *J. Mol. Biol.* **236**, 556–571.
31. Loeb, D. D., Swanstrom, R., Everitt, L., Manchester, M., Stamper, S. E. & Hutchison, C. A., Jr (1989). Complete mutagenesis of the HIV-1 protease. *Nature*, **340**, 397–400.
32. Axe, D. D., Foster, N. W. & Fersht, A. R. (1998). A search for single substitutions that eliminate enzymatic function in a bacterial ribonuclease. *Biochemistry*, **37**, 7157–7166.
33. Eschenburg, S., Genov, N., Peters, K., Fittkau, S., Stoeva, S., Wilson, K. S. & Betzel, C. (1998). Crystal structure of subtilisin DY, a random mutant of subtilisin Carlsberg. *Eur. J. Biochem.* **257**, 309–318.
34. Alexander, P. A., Ruan, B. & Bryan, P. N. (2001). Cation-dependent stability of subtilisin. *Biochemistry*, **40**, 10634–10639.
35. Zhang, F., Strand, A., Robbins, D., Cobb, M. H. & Goldsmith, E. J. (1994). Atomic structure of the MAP kinase ERK2 at 2.3 Å resolution. *Nature*, **367**, 704–711.
36. Harel, M., Kryger, G., Rosenberry, T. L., Mallender, W. D., Lewis, T., Fletcher, R. J. *et al*. (2000). Three-dimensional structures of *Drosophila melanogaster* acetylcholinesterase and of its complexes with two potent inhibitors. *Protein Sci.* **9**, 1063–1072.
37. Sussman, J. L., Harel, M., Frolow, F., Oefner, C., Goldman, A., Toker, L. & Silman, I. (1991). Atomic structure of acetylcholinesterase from *Torpedo californica*: a prototypic acetylcholine-binding protein. *Science*, **253**, 872–879.
38. Bourne, Y., Taylor, P. & Marchot, P. (1995). Acetylcholinesterase inhibition by fasciculin: crystal structure of the complex. *Cell*, **83**, 503–512.
39. Kryger, G., Harel, M., Giles, K., Toker, L., Velan, B., Lazar, A. *et al*. (2000). Structures of recombinant native and E202Q mutant human acetylcholinesterase complexed with the snake-venom toxin fasciculin-II. *Acta Crystallog. D: Biol. Crystallog.* **56**, 1385–1394.
40. Helmreich, E. & Cori, C. F. (1964). The role of adenylic acid in the activation of phosphorylase. *Proc. Natl Acad. Sci. USA*, **51**, 131–138.
41. Barford, D., Hu, S. H. & Johnson, L. N. (1991). Structural mechanism for glycogen phosphorylase control by phosphorylation and AMP. *J. Mol. Biol.* **218**, 233–260.
42. Oikonomakos, N. G., Skamnaki, V. T., Tsitsanou, K. E., Gavalas, N. G. & Johnson, L. N. (2000). A new allosteric site in glycogen phosphorylase b as a target for drug interactions. *Struct. Fold. Des.* **8**, 575–584.
43. Barford, D. & Johnson, L. N. (1989). The allosteric transition of glycogen phosphorylase. *Nature*, **340**, 609–616.
44. Cobb, M. H., Boulton, T. G. & Robbins, D. J. (1991). Extracellular signal-regulated kinases: ERKs in progress. *Cell Reg.* **2**, 965–978.
45. Boulton, T. G., Yancopoulos, G. D., Gregory, J. S., Slaughter, C., Moomaw, C., Hsu, J. & Cobb, M. H. (1990). An insulin-stimulated protein kinase similar to yeast kinases involved in cell cycle control. *Science*, **249**, 64–67.
46. Anderson, N. G., Maller, J. L., Tonks, N. K. & Sturgill, T. W. (1990). Requirement for integration of signals from two distinct phosphorylation pathways for activation of MAP kinase. *Nature*, **343**, 651–653.
47. Crews, C. M., Alessandrini, A. & Erikson, R. L. (1992). The primary structure of MEK, a protein kinase that phosphorylates the ERK gene product. *Science*, **258**, 478–480.
48. Ahn, N. G., Seger, R., Bratlien, R. L., Diltz, C. D., Tonks, N. K. & Krebs, E. G. (1991). Multiple components in an epidermal growth factor-stimulated protein kinase cascade. *In vitro* activation of a myelin basic protein/microtubule-associated protein 2 kinase. *J. Biol. Chem.* **266**, 4220–4227.
49. Canagarajah, B. J., Khokhlatchev, A., Cobb, M. H. & Goldsmith, E. J. (1997). Activation mechanism of the MAP kinase ERK2 by dual phosphorylation. *Cell*, **90**, 859–869.
50. Casari, G., Sander, C. & Valencia, A. (1995). A method to predict functional residues in proteins. *Nature Struct. Biol.* **2**, 171–178.
51. Laskowski, R. A., Luscombe, N. M., Swindells, M. B. & Thornton, J. M. (1996). Protein clefts in molecular recognition and function. *Protein Sci.* **5**, 2438–2452.
52. Zvelebil, M. J. & Sternberg, M. J. (1988). Analysis and prediction of the location of catalytic residues in enzymes. *Protein Eng.* **2**, 127–138.
53. Siew, N. & Fischer, D. (2003). Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins: Struct. Funct. Genet.* **53**, 241–251.
54. Siew, N., Azaria, Y. & Fischer, D. (2004). The ORFanage: an ORFan database. *Nucl. Acids Res.* **32**, D281–D283.
55. Axe, D. D. (2000). Extreme functional sensitivity to conservative amino acid changes on enzyme exteriors. *J. Mol. Biol.* **301**, 585–595.
56. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al*. (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
57. Bairoch, A. (1993). The ENZYME data bank. *Nucl. Acids Res.* **21**, 3155–3156.

58. Hubbard, S. J. (1996). *Naccess* (2.1.1 edit.), Biomolecular Structure and Modelling Unit, University College, London, UK.
59. Mizuguchi, K., Deane, C. M., Blundell, T. L. & Overington, J. P. (1998). HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* **7**, 2469–2471.
60. Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E. & Ben-Tal, N. (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.
61. Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Netw.* **1**, 215–239.
62. Beauchamp, M. A. (1965). An improved index of centrality. *Behav. Sci.* **10**, 161–163.
63. Sabidussi, G. (1966). The centrality of a graph. *Psychometrika*, **31**, 581–603.
64. Liang, J., Edelsbrunner, H. & Woodward, C. (1998). Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* **7**, 1884–1897.
65. Noguchi, T. & Akiyama, Y. (2003). PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. *Nucl. Acids Res.* **31**, 492–493.
66. Henikoff, S., Henikoff, J. G., Alford, W. J. & Pietrokovski, S. (1995). Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene*, **163**, GC17–GC26.

Edited by B. Honig

(Received 15 June 2004; received in revised form 8 October 2004; accepted 19 October 2004)