doi:10.1016/j.jmb.2004.08.022 J. Mol. Biol. (2004) 342, 1487–1504

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Distinguishing Structural and Functional Restraints in Evolution in Order to Identify Interaction Sites

Vijayalakshmi Chelliah, Lan Chen, Tom L. Blundell and Simon C. Lovell*

Department of Biochemistry
University of Cambridge
80 Tennis Court Road
Cambridge CB2 1GA, UK

Structural genomics projects are producing many three-dimensional structures of proteins that have been identified only from their gene sequences. It is therefore important to develop computational methods that will predict sites involved in productive intermolecular interactions that might give clues about functions. Techniques based on evolutionary conservation of amino acids have the advantage over physiochemical methods in that they are more general. However, the majority of techniques neither use all available structural and sequence information, nor are able to distinguish between evolutionary restraints that arise from the need to maintain structure and those that arise from function. Three methods to identify evolutionary restraints on protein sequence and structure are described here. The first identifies those residues that have a higher degree of conservation than expected: this is achieved by comparing for each amino acid position the sequence conservation observed in the homologous family of proteins with the degree of conservation predicted on the basis of amino acid type and local environment. The second uses information theory to identify those positions where environment-specific substitution tables make poor predictions of the overall amino acid substitution pattern. The third method identifies those residues that have highly conserved positions when three-dimensional structures of proteins in a homologous family are superposed. The scores derived from these methods are mapped onto the protein three-dimensional structures and contoured, allowing identification clusters of residues with strong evolutionary restraints that are sites of interaction in proteins involved in a variety of functions. Our method differs from other published techniques by making use of structural information to identify restraints that arise from the structure of the protein and differentiating these restraints from others that derive from intermolecular interactions that mediate functions in the whole organism.

Keywords: protein function; functional sites; evolution; environment-specific substitution tables

*Corresponding author

Introduction

Although structural studies of proteins have been traditionally carried out on systems that have been functionally well-characterized, structural genomic projects are reversing this tendency, in that they explicitly select proteins that are unrelated to others of known structure and function.[1–3] As a consequence, the structures of many proteins that are poorly characterized in terms of function and biochemistry are now being determined. In order to make best use of these structures, it is essential to identify their functions.

The function of a protein may be described on any of several levels. For example the molecular function of an enzyme depends on its specificity

Present address: S. C. Lovell, School of Biological Sciences University of Manchester, Smith Building, Oxford Road, Manchester M13 9PT, UK.

Abbreviations used: pdb, Protein Data Bank; RMSD, root-mean-square deviation; LDH, lactate dehydrogenase; MDH, malate dehydrogenase; IGF, insulin-like growth factor.

E-mail address of the corresponding author: simon.lovell@man.ac.uk

and the reaction it catalyses; the cellular function may depend additionally on its temporal and spatial expression in the cell, and the physiological function on the organ in which the expressing cells are found (for a review see van Helden *et al.*[4]). As we move up this hierarchy, additional information is required to understand the function. However, each level of function depends on the level beneath. Thus, assignment of molecular function, which is often defined in terms of the productive interactions the protein makes in the cell, is an essential first step. Localisation of such interactions to so-called "functional sites" or "interaction sites" will allow us to understand how the protein may recognise other molecules, to gain clues about its likely function at the level of the cell and organism and to identify important binding sites that may serve as useful targets for pharmaceutical design.

There have been a wide variety of approaches to the problem of functional site detection. Functional sites have been assigned on the basis of changed binding or enzymatic properties resulting from chemical modification of specific amino acid residues or site-directed mutagenesis.[5,6] Sequence motif databases, such as PROSITE,[7] which identify sequentially conserved residues and also depend on literature searches, record specific residues likely to be involved in function. This information has been used to annotate various sequence alignment databases, such as Pfam[8] and HOMSTRAD[9]†.

Three-dimensional descriptors of functional sites have an advantage over linear descriptors as the sites themselves are usually comprised of discontinuous regions of protein sequence, giving rise to weak linear sequence motifs but stronger three-dimensional motifs. Kasuya & Thornton[10] have correlated one-dimensional PROSITE motifs with tertiary structures to define three-dimensional functional patterns. Skolnick and co-workers have annotated protein structures with experimentally derived functional information to produce what they term a "fuzzy functional form", which is a three-dimensional descriptor of the functional site of a protein.[11–15] Since these annotations are derived from experimentally determined data they are likely to be of high quality, but are laboriously gathered. There have been, therefore, several attempts to identify or predict functional/interaction sites computationally. These studies fall into two broad classes: those that use physical features of the protein and those that seek to identify evolutionary conservation.

Several authors have analysed protein structures in search of steric strain or other types of high-energy conformations. Hertzberg & Moult[16] found that Ramachandran outliers are often correlated with functional sites. Similarly, Heringa & Argos[17] found that non-rotameric side-chains forming interacting clusters are often found in functional sites. Elcock[18] extended this work by predicting functional sites on the basis that they have a high energy, as calculated from a molecular mechanics force field. Ota *et al.*[19] have used a similar method in combination with identification of conserved residues to predict catalytic residues in enzymes.

Laskowski *et al.*[20] analysed clefts in proteins and found the largest cleft tends to be involved in ligand binding areas. The same group has also characterised protein–protein interaction sites by the analysis of the nature of interface residues compared with other residues on the protein surface,[21] finding that in homo-complexes they are significantly more hydrophobic than the rest of the protein surface. Functional sites have been identified on the basis of clusters of charge,[22] particularly mixed charge residues[23] and clusters of conserved, exposed, polar residues.[24]

A marked disadvantage of identifying interaction/functional sites from physical chemistry is that different functional sites have different characteristics. For example, nucleic acid binding sites on proteins can be identified by their positively charged nature,[25] but this is not generally applicable to protein–protein interaction sites. Indeed, protein–protein interaction interfaces in homo and hetero-complexes are significantly different.[21] Further differences will occur in systems where assembly and disassembly is critical to function, and the protein–protein interactions are not enduring. In contrast, with the notable exception of immunoglobulins, which have their binding site produced by recombination of VDJ genes, almost all protein functional sites are optimised through mutation and Darwinian selection. Mutations that change or abolish function will usually be directly selected against and hence functional sites will be the most highly conserved regions of a protein. It has long been known that residue identity is highly conserved in enzyme active sites,[26] although there are more exceptions to this than was originally envisaged.[27,28] Attempts to produce a general method for identifying functional sites have centred on conservation of sequence. However, conservation of structure can also be used,[29] or, as here a combination of sequence and structure conservation.

The observation that those regions in which main-chain conformation is most conserved in enzymes is likely to be the active site was first made by Chothia & Lesk.[30] McPhalen *et al.*[31] developed an iterative three-dimensional fitting method to define structurally conserved regions in proteins. This was applied by Irving *et al.*[29] to identify enzyme active sites in homologous families.

The most widely used method based on evolutionary conservation of sequence is "evolutionary trace".[32–35] In this technique a phylogenetic tree is constructed based on sequences, and the tree partitioned. Residues that are conserved in different partitions, or in all partitions, are highlighted on the structure, which is then examined visually. The technique has had a

† http://www-cryst.bioc.cam.ac.uk/homstrad

number of successes (for a review, see Litcharge & Sowa[36]).

The original method relied on visual identification of clusters making the technique difficult to apply automatically to the whole database of structures, although subsequent addition of automated clustering has solved this problem.[34] Other authors have removed the dependence on absolute conservation, using a substitution table to allow for substitution of physiochemically similar residues,[37,38] and to allow for non-uniform rates of evolution at each site. Landgraf *et al*.[39] have identified sequence conservation in three-dimensional clusters, and found it adds sensitivity to the evolutionary trace method.

There are, however, outstanding problems. Bork & Koonin[40] and Karp[41] discuss the problems with assigning function from the literature. Annotation of databases with experimental information suffers from the rapid growth of sequence and structural information that is not matched by the growth of experimental information, a gap that is growing. Annotation of sequence databases has the additional problem of the linear form of the data representation not always being compatible with the three-dimensional form of the functional site.

Predictions based on description of the physical characteristics of functional sites usually work well for the class of sites studied (usually enzymes) but often cannot be generalised. Predictions based on high-energy or strained conformations will often find active sites, but are unlikely to be able to identify many other classes of functional sites, particularly those involved in protein–protein interactions.

Methods that identify sequence conservation do not distinguish evolutionary restraints that arise from function from those that arise from structure, and the consideration of chemical similarity does little to make the distinction. Because the core of the protein is likely to be conserved for structural reasons, often only surface residues are analysed, simply by viewing a space-filling representation of the protein, or a solvent accessible surface. However, solvent accessible surfaces miss many residues that provide hydrogen bonds or that become accessible after conformational changes. This is a particular problem for catalytic residues of enzymes, which can often be relatively inaccessible to solvent.

The conservation of amino acid residues has been shown to be strongly dependent on the environment in which they occur in the folded protein and amino acid substitution tables that give the likely substitutions of amino acids in particular local environments have been derived.[42,43] We present here a method for using these environment-specific substitution tables to distinguish those restraints placed on protein structure from additional restraints due to particular functions mediated by interactions with other molecules. We find that the clusters of residues apparently subjected to these additional restraints in evolution correlate well with the functional sites in proteins defined by experimental methods. We also analyse conservation of local structure in homologous families of proteins and develop a term to describe structural conservation that can be used to increase the accuracy of functional site identification. The method relies on the clustering of residues in three-dimensional space.[34] We apply it to a set of well-characterised protein families and are able to identify functional sites. The technique is fast, automatic and predicts functional sites with a high degree of accuracy.

Results

Datasets used

We have selected three sets of families from the HOMSTRAD database. HOMSTRAD was chosen because it provides structure-based sequence alignments of evolutionarily related protein families that can be used as the basis for collecting sequence homologues. The three datasets were: a "jack-knife" set of ten families that were not used to derive the original substitution tables, a set of enzymes that were added to HOMSTRAD after we derived the substitution tables and performed the initial analysis which we term the "new" set, and a "non-independent" set of enzyme families which were in the set of proteins used to derive the substitution tables, but which are included for completeness. The jack-knife set consists of ten families containing 136 structures, the new set consists of 154 families containing 392 structures, and the non-independent set consists of 80 families containing 346 structures.

The jack-knife set is detailed in Table 1. These families were selected on the basis of availability of extensive information in the literature about their active sites and structure, and at least one protein–substrate or protein–substrate analogue structure available which has the same binding mode as the substrate. Active site residues of each family, also detailed in Table 1, were identified using: (i) the literature; (ii) PROSITE[44–46] sequence motifs; (iii) complexes with ligands in PDB;[47] and (iv) "ACT_SITE" records in the SwissProt database.[48] Both catalytic and ligand-binding residues are included. These families have been studied widely and characterised extensively, so the literature information is comprehensive. On average, we have identified 10.9 "functional" residues per family. The average number of proteins with known structure in each family is high (13.6), and so the structural divergence information is extensive. We consider data derived from this set to be most reliable.

The set of new families are statistically independent of the set used to derive the substitution tables as they were added to HOMSTRAD after their calculation and our initial analysis. Information about functional and catalytic residues was derived from the literature. Because of lack of

Table 1. Results of HOMBLAST and FUGUESEQ for the ten families

Family name	No. of structures	No. of sequences	Example structure	Active site residues
Serine proteinases-trypsin	27	187	1kig	57, 102, 143, 146, 147, 192, 195, 216, 224, 226
Cysteine proteinases	13		1mem	9, 23, 25, 26, 66, 67, 159, 160, 177
Aspartic proteinases	13	80	1smr	30, 32, 34, 35, 73, 75, 112, 120, 215, 217, 218, 219
Matrix metalloproteinases	6	67	1bqq	186, 188, 201, 214, 239, 240, 242, 243, 249, 259
Lactate/malate dehydrogenase	14	112	2 cmd	81, 87, 119, 149, 153, 177, 210, 223
Glutatione *S*-transferase	14	72	1gsu	6, 7, 9, 11, 12, 45, 49, 58, 59, 71, 72, 115, 208, 209
Serine/threonine protein kinases	15	253	1phk	25, 31, 48, 104, 106, 110, 149, 154, 151, 186
PLA-2 phospholipase	18	80	1pob	2, 5, 9, 28, 29, 31, 44, 47, 48, 51, 63, 93, 100
Triose phosphate isomerase	10	107	1hti	11, 13, 95, 96, 97, 165, 209, 210, 230, 232
Xylose isomerase-like	6	200	4xia	15, 53, 56, 136, 180, 182, 214, 216, 219, 245, 254, 256, 292

The number of structures available in HOMSTRAD, number of clusters ($\geq$80% identical within the members of each cluster, and not more than 80% identical between the members of each cluster), and the pdb code of the representative structure in HOMSTRAD for all the ten families and the residues considered to be functional.

detailed information about protein–ligand complexes and lack of knowledge about relevant binding modes, we did not use this information to determine which residues are in contact with substrates. The average number of residues classed as "functional" is 4.8. This is much lower than the jack-knife set, we assume because of the lack of binding information. The average number of structures per family is 2.5, and therefore structural divergence information is less reliable.

The non-independent set was included to increase the size of the test sets. Its characteristics are similar to the new set. The average number of structures per family is 4.3 and the average number of residues counted as functional is 4.7. It should be noted that this set is not statistically independent, and therefore we consider data derived from this set the least reliable.

In order to allow comparisons we also identified those residues from the jack-knife set which could only be identified from the literature and other databases, in the same manner as for the new and non-independent sets. This "literature only" annotation resulted in an average of 3.9 functional residues per protein. If the results from the jack-knife set differed from those derived from the other datasets, this information allows us to identify whether this is due to differences in the datasets, or differences in the amount of information we have about those datasets.

From these dataset families we attempted to identify evolutionary restraints on sequence and structure. We have formulated three scoring systems: the first is based on identifying sequence conservation above that predicted from environment-specific substitution tables; the second identifies where environment-specific substitution tables make poor predictions of the overall sequence distribution at each alignment position; the third identifies where main-chain conformation is highly conserved.

Sequence-based scoring systems

Two of the scoring systems are sequence-based. Both of them compare the observed substitution pattern to the expected substitution pattern predicted from the environment-specific substitution tables. It is necessary to use environment-specific tables because the degree of conservation for an amino acid depends both on its type and its local environment.[42,43]

The substitution tables are calculated from the whole of the HOMSTRAD database, and therefore represent the average substitution pattern for an amino acid residue of a given type in a particular environment. Both scoring systems attempt to find unusual residues; the divergence score finds those that generally have atypical substitution patterns, whereas the conservation score finds those that are unusually well conserved for their type and environment. The observed substitution pattern was calculated from the HOMSTRAD family and the sequence homologues collected by running PSI-BLAST.[49]

Residues may have atypical substitution patterns for a number of reasons. There may be deficiencies in the substitution tables, or it may be inappropriate to assign environments to homologous proteins of unknown structure. These will lead to noise in the method. There may also be errors in the multiple sequence alignment, particularly when using automatic methods.

Most unusual substitution patterns will arise because the environment has been incompletely or inappropriately assigned, and this may give clues to

binding sites. Very often the oligomerisation state of a protein is unknown or incorrectly defined. For example, if the model used is a monomer, whereas in the organism the protein exists as a dimer, the substitution pattern would be predicted to be exposed to solvent but the observed evolutionary restraints would be of a buried residue. An example is the protomer of malate dehydrogenase. If either of the sequence-based scores is calculated, residues in the homodimer interface score highly in addition to the substrate and cofactor binding sites. This is because residues that are buried in the dimer are inappropriately assigned a solvent-exposed environment too. If the solvent accessibility is calculated for the dimer, the substitution pattern is well predicted. Thus, the scoring systems highlight those aspects of the system that are not adequately described by the input.

However, most high scores in the enzymes we have considered will arise because the protein interacts with other molecules to achieve catalysis. If the function is conserved amongst the proteins considered, there will be strong evolutionary pressure for the residues involved in protein interactions that mediate the function to remain unchanged or conservatively varied. This will be true of cofactor binding, enzyme–substrate interactions, protein interactions in multi-protein complexes, allosteric effectors and so on.

Figure 1 shows that applications of our method using the two sequence-scoring schemes are generally similar, with a correlation coefficient of 0.68. Differences are mostly seen when residues are deleted in other family members. This is due to the different treatment of deletions by the two scoring schemes. As neither scoring system leads to these residues having high scores, the difference is unimportant when identifying functional sites.

Figure 2 shows the distribution of *Z*-scores for the

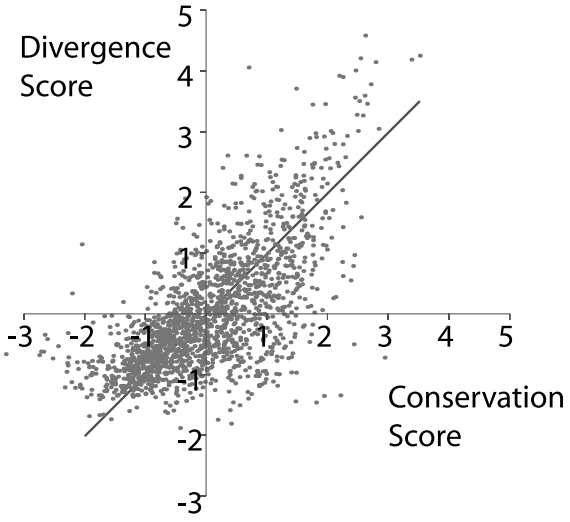

Figure 1. Comparison of conservation and divergence scores for the ten families. The line indicates equality between the two scores.

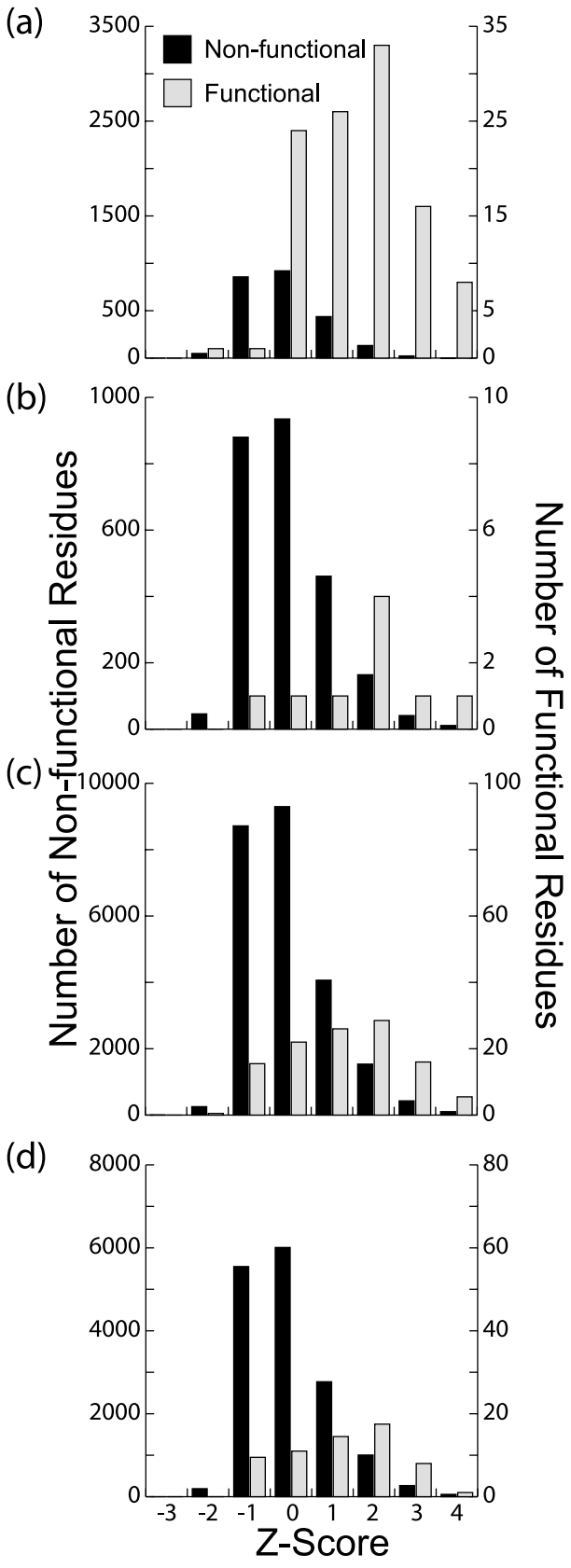

Figure 2. Distribution of *Z*-scores for proteins in the various data sets. Light bars and the left hand axis indicate functional residues, grey bars and the right hand axis indicate non-functional residues. (a) Jack-knife set; (b) jack-knife set with literature annotation of functional residues only; (c) new set; (d) non-independent set.

enzymes in the three test sets. Active site residues are highlighted. Active site residues are predominantly the highest scoring residues in the set. The residues described as functional are those that are described in the literature as functional (all sets) or make direct contact with a substrate or substrate analogue (jack-knife set only; Figure 2(a)). As a control we also show residues for the jack-knife set with residues annotated from the literature and databases only (Figure 2(b)). Both the number and distribution of scores for the functional residues are similar for the three datasets when the same criteria are used for defining functional residues. However, if ligand-binding information (not available for the new and non-independent datasets) is not used, the majority of functional residues are missed; this, then, is a deficiency in the level of annotation rather than our methods.

Unusual evolutionary restraints may act upon both those residues directly involved in function (for example catalytic or ligand-binding residues), or those that have a less direct role. In the aspartic proteinase 1smr, the catalytic residues are Asp32 and Asp215, both of which have conservation scores of 1.9. Thr33 (and similarly Thr216) has a conservation score of 1.8 (divergence scores of 0.8 and −0.1, respectively, due to their buried nature; see below). Threonine 33 makes side-chain to main-chain hydrogen bonds to Val214 and Thr216. Thus, the residue neighbouring one catalytic residue makes direct interactions to the two residues flanking another catalytic residue, almost certainly holding the two catalytic residues in the correct relative orientation. It is conserved, and so has a high conservation score. Because Thr33 is not directly involved in catalysis or ligand binding it would not traditionally be counted as a catalytic residue. Only careful analysis of the structures identifies its role in stabilising the active site through the so-called "fireman's grip".[50]

Structure-based scoring schemes

It has been suggested by Chothia & Lesk[30] that the functional site of a protein should be the most conserved structurally. Irving *et al*.[29] predicted the active site of YabJ from *Bactilis subtilis* and YjgF from *Escherichia coli* using this approach. We have developed a scoring system for describing the degree of structural conservation in proteins where multiple structures are available. We find that the set of equivalences used for superposition influences the final score, and so we choose our initial equivalence set based on the residues with highest conservation score, which is then expanded to include the most structurally conserved residues up to a cut off of 1.5 Å root-mean-square deviation (RMSD). It is critically important that only those structures in the same liganded state (all liganded or all unliganded) are included, otherwise ligand-induced conformational changes become confused with structural divergence.

The structure-based score is poorly correlated with the sequence-based scores (Figure 3), with correlation coefficients of 0.35 (structure *versus* divergence) and 0.40 (structure *versus* conservation). This is probably because the structure-based score is very sensitive to small conformational differences that arise from crystal packing, varying crystallisation conditions, and different ligands as well as errors in the crystal structure analysis. Sequence conservation methods are less susceptible to errors. For this reason, we have combined them using empirically determined weights of 0.35 (structure) and 0.65 (sequence).

The effect of considering amino acid environment

The effect of environment-specific substitution tables can be seen if we compare sequence

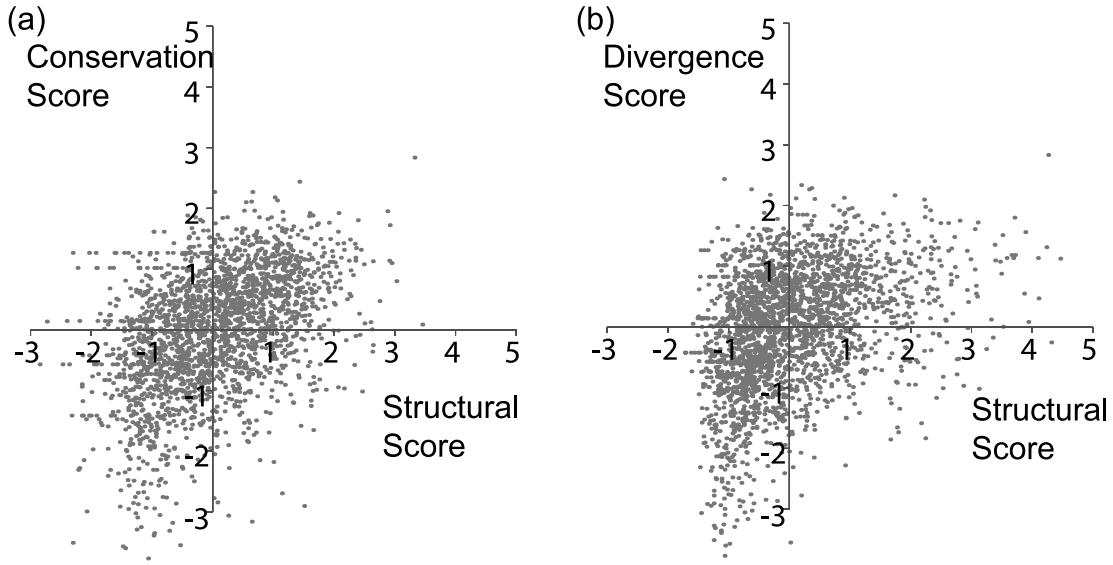

Figure 3. The relationship between the structural and sequence-based scores.

Table 2. Residues whose scores are most effected by the use of environment-specific substitution tables

Conservation score			Divergence score		
Score difference	Residue	Role in protein	Score difference	Residue	Role in protein
Most strongly down-weighted residues					
−0.48	Tyr129	Buried, H-bonding	−1.73	Asp200	Buried charge
−0.43	Gln114	Charged, H-bonding	−1.57	Thr268	
−0.42	Asp200	Buried charge	−1.37	His123	Buried charge
−0.38	Thr268		−1.34	Tyr129	Buried, H-bonding
Most strongly up-weighted residues					
0.3	Ala203		0.99	Met87	ATP binding
0.31	Lys133	Mg/ATP binding	1.04	Gly180	In active site
0.34	Leu138	ATP binding	1.12	Leu138	ATP binding
0.44	Ala39	ATP binding	1.34	Ala39	ATP binding

conservation scores using standard and environment-specific tables. The difference between these two scores is shown in Table 2, for the lactate/malate dehydrogenase family. Positive values are those that are up-weighted by the environment-specific tables, negative values are down-weighted. The four most strongly up-weighted and four most strongly down-weighted residues are given. As can be seen, those residues that are functional are up-weighted, whereas those that are likely to be conserved for structural reasons[42,43] are down-weighted.

Some active site residues have a relatively low divergence score. The residues that form the catalytic triad of serine proteinases are His57, Asp102 and Ser195. His57 and Ser195 have divergence *Z*-scores of 3.3 and 4.3, respectively, and are predicted to be functional. Asp102 has a *Z*-score of 0.5, normally indicating a non-functional residue. Asp102 is probably charged, hydrogen-bonded and in a buried environment. This is a very highly conserved residue/environment type.[42] As it is also expected to be highly conserved for structural reasons, its substitution pattern is not changed significantly by the further restraint of function. Thus, it is not well predicted by the substitution tables and it has a low score. The other two members of the catalytic triad (His57 and Ser195) are solvent-accessible and not expected to be conserved: as a consequence these score highly. Similarly, Asp150 of malate dehydrogenase (pdb code 2 cmd) and Asp93 of phospholipase (pdb code 1pob) are catalytic but are in a buried environment and are therefore not well predicted to be functional by the divergence score for the same reason.

If the divergence score is calculated using non-environment-specific substitution tables, all of these catalytic residues score highly. However, this results in a large number of other residues additionally having high scores. These other residues are in those conserved for structural reasons, and so the use of non-environment-specific tables removes one of the main strengths of our method.

The conservation score has different characteristics. It does not correct so effectively for the environmental effects; this can be seen in Table 2. The score difference between environment-specific and non-environment-specific Table is two to three times larger for the divergence score compared with the conservation score. In most cases, therefore, the divergence score will be preferable as it contains more information about the environment. The disadvantage of down-weighting buried, charged, catalytic residues is, however, only a feature of the divergence score and not the conservation score. Where such buried, charged functional residues are suspected (for example when identifying active sites) it is preferable to use both scoring systems, paying close attention to those residues that score differently when the two methods are compared.

Success in prediction

We investigated whether we could use these evolutionary restraints to identify functional sites in

Table 3. Atoms used to position score in three dimensions before contouring

Residue	Atoms defined as functional
Ala	C^{β}
Arg	$N^{\varepsilon 1}$, C^{ζ}, $N^{\eta 1}$, $N^{\eta 2}$
Asp	$O^{\delta 1}$, $O^{\delta 2}$
Asn	$O^{\delta 1}$, $N^{\delta 2}$
Cys	S^{γ}
Gln	$O^{\varepsilon 1}$, $N^{\varepsilon 2}$
Glu	$O^{\varepsilon 1}$, $O^{\varepsilon 2}$
Gly	C^{α}
His	$C^{\delta 2}$, $C^{\varepsilon 1}$
Ile	$C^{\gamma 1}$, $C^{\gamma 2}$, $C^{\delta 1}$
Leu	$C^{\delta 1}$, $C^{\delta 2}$
Lys	N^{ζ}
Met	S^{δ}, C^{ε}
Phe	C^{γ}, $C^{\delta 1}$, $C^{\delta 2}$, $C^{\varepsilon 1}$, $C^{\varepsilon 2}$, C^{ζ}
Pro	C^{β}, C^{γ}, C^{δ}
Ser	O^{γ}
Thr	O^{γ}
Trp	$N^{\varepsilon 1}$
Tyr	C^{γ}, $C^{\delta 1}$, $C^{\delta 2}$, $C^{\varepsilon 1}$, $C^{\varepsilon 2}$, C^{ζ}, O^{η}
Val	$C^{\gamma 1}$, $C^{\gamma 2}$

Where more than one atom is listed the average position is used.

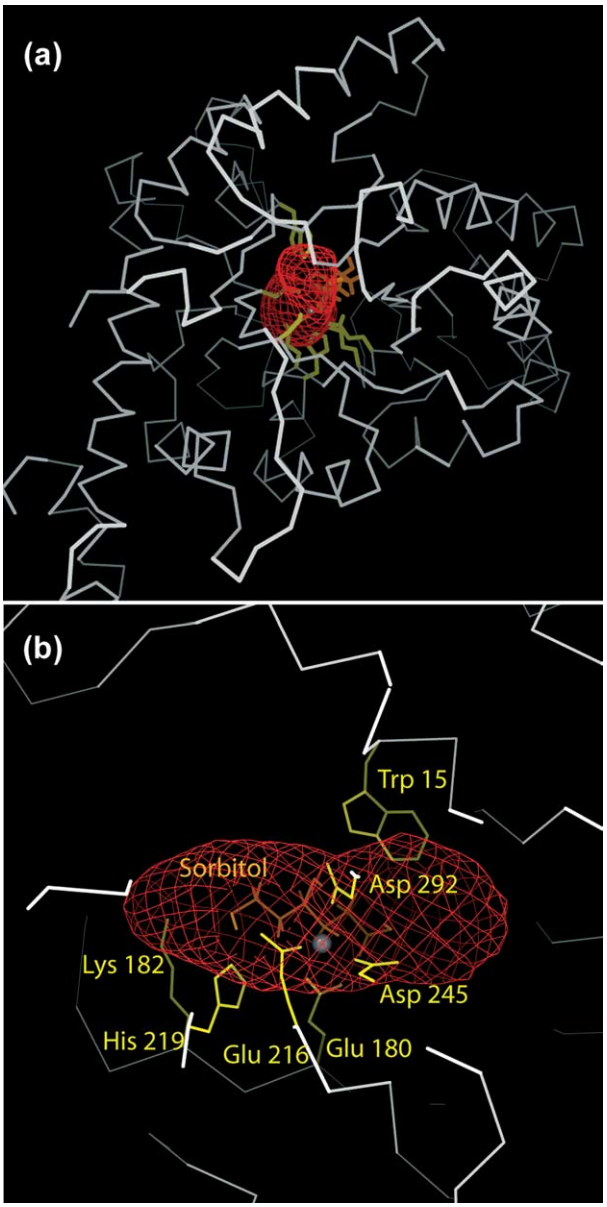

Figure 4. Contouring of the combined score at the 0.01 score $Å^{-3}$ level for xylose isomerase (4xia) with active site residues shown. (b) A closer view of the active site with active site residues and the sound inhibitor sorbitol shown.

proteins in the test sets. In addition to having high scores, functional residues should cluster in three dimensions to form one or more functional sites. In the case of our test set one cluster of high-scoring residues should be the active site of the enzyme; the appearance of specificity sub-sites, oligomerisation or other functional sites will depend on how extensively these features are shared by the enzymes whose sequences and structures are conserved (see below).

We have mapped the score onto the protein structure. The coordinate chosen in each case is that of the atom most likely to be involved in function and is given in Table 3. The scores were smoothed, converted into a grid of conservation density and this grid contoured. An example is shown in Figure 4.

In order to test whether the correct site is being identified, the highest grid point was determined and compared to the centre of the active site. The active site centre was defined as the average position of functional atoms in all active site residues. The distance between the highest grid point and the active site centre is shown in Table 4. Generally the two sequence-based methods are equally useful, and the structure-based score slightly less so. The combination of divergence and structural scores gives the best results. Data on the distance between actual and predicted functional sites for all data sets are given in Figure 5.

The number of active site residues within the largest high-scoring cluster of grid-points (see Methods) is given in Table 5. We are able to identify successfully a large proportion of active site residues. Because all sequences are included, we are able to highlight only those residues that are functional (and therefore have unusual evolutionary restraints) across the entire family. Those residues that determine specificity will only score highly if the sequences are separated into a group of differing specificity (see below).

Conformational change

In order for the technique to be useful in identifying functional sites, it is important that it

Table 4. Distance from centre of active site to peak of conservation density for the jack-knife set

		Distance (Å)			
Family	pdb code	Cons.	Diverge.	Struc.	Comb.
asp	1smr	4.1	3.5	5.4	4.1
kinase	1phk	3.2	5.9	2.5	3.0
cys	1mem	3.8	5.8	4.3	4.1
ldh	2 cmd	4.2	3.0	4.9	3.8
mmp	1bqq	3.1	6.7	7.6	2.1
gluts	1gsu	4.2	9.4	5.0	3.9
phoslip	1pob	3.7	3.1	5.5	3.6
sermam	1kig	2.4	5.9	3.4	2.1
tim	1hti	2.1	2.4	3.6	1.8
xia	4xia	2.1	3.2	3.8	2.2

Cons., conservation score alone; Diverge., divergence score alone; Struc., structure-based score alone; Comb., combination of divergence and structure-based score.

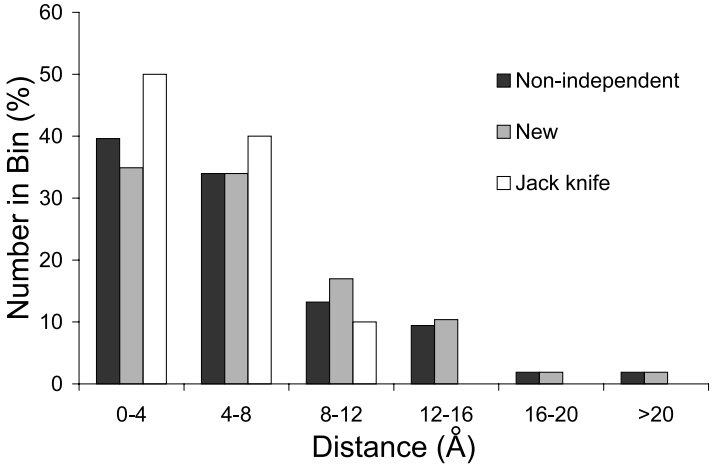

Figure 5. The distance between the predicted and actual functional site for proteins in the three data sets.

not be over-sensitive to ligand-induced conformational change. We investigated the effect of conformational change by examining the lactate/malate dehydrogenase family. In malate dehydrogenase there is a large conformational change in the active site loop upon substrate binding.[51] This is seen clearly when binary (4mdh, with co-factor: NAD) and ternary (5mdh, with co-factor: tetrahydroNAD and substrate: α-ketomalonate) structures are compared. Arginine 97 undergoes particularly large movements. In the binary complex Arg97 is found to protrude into the solvent and the active site loop adopts an "open" conformation. This contrasts with the ternary structure where Arg97 is involved with substrate binding and is pointing into the active site. The Arg97 is solvent accessible in the binary complex and almost buried in the ternary complex, having only 5.9% of its surface exposed. When Arg97 is in the buried environment (ternary complex) the high degree of conservation is expected due structural as well as functional considerations. Thus, the pattern of substitution is well predicted by the substitution tables and the residue has a low divergence score of −0.2.

Of the family of 15 structures, Arg97 is in a buried environment in only three. We find that when the substitution pattern of each of the known structure was calculated from the environment-specific substitution table and averaged over all structures in order to find the predicted substitution pattern, this position has a divergence score of 2.3 and the residue is predicted to be functional. Of course the prediction is even better when only those residues in an open, binary structure are considered.

The ligand-induced conformational change makes a large difference to the score for a specific residue. Even though we would advise as a consequence grouping structures in the same liganded state together, even for the sequence based methods, when the scores are mapped onto the structure and smoothed, there is much less difference. The maximum in the conservation density moves by only 1.3 Å when the binary and ternary complexes are considered. The only residue that changes from being inside to outside the largest high-scoring cluster is Arg97 itself. Our sequence-based techniques are, therefore, relatively insensitive to ligand-induced conformational change, even in this rather extreme case.

Table 5. Numbers of residues predicted to be functional, as compared to those describe as functional in the literature, for all families in the jack-knife set and all data sets

	Example	Correct	False pos.	Missed
A. *Family*				
Aspartic proteases	1smr	7	1	5
Glutathione S transferase	1gsu	3	4	10
Cysteine proteases	1mem	5	3	4
Ser/Thr protein kinase	1phk	4	3	6
Lactate/malate dehydrogenase	2 cmd	4	2	4
Metalloproteases	1bqq	4	2	6
Phospholipase	1pob	5	1	8
Serine proteases	1kig	3	6	7
Triose phosphate isomerase	1hti	6	1	5
Xylose isomerase	4xia	4	6	9
B. *Data set*				
Jack-knife		45	29	64
Jack-knife, lit. info. only		26	46	13
New		295	860	444
Non-independent		156	430	218

Effect of inclusion of different proteins

Initially we included all homologous sequences (as collected using PSI-BLAST) in our calculation of observed substitution patterns. This is a simple, automated way of collecting sequence homologues, and is successful at identifying functional sites. It has the feature, however, of highlighting only those residues that are common across the entire set of sequences. This means we can easily identify catalytic residues, but are less successful in identifying specificity-determining residues, as the latter, by their nature will differ across the homologous family.

If it is desirable to identify specificity-determining residues, the sequences can be divided into sub-groups and the observed substitution pattern recalculated. We have analysed the lactate/malate dehydrogenase family to see the effect of using single-specificity sub-groups.

The lactate/malate dehydrogenase family can clearly first be subdivided into the lactate dehydrogenases (LDHs) and the malate dehydrogenases (MDHs). The mammalian lactate dehydrogenases can then be further divided into liver, heart and muscle forms,[52] whereas the malate dehydrogenases can be further divided into the cytosolic and mitochondrial forms. The MDHs are usually dimeric, whereas the LDHs are usually tetrameric. The MDH dimerisation interface and one of the LDH interfaces (here termed the LDH dimerisation interface) are generated by equivalent symmetry. The LDHs additionally have another interface which has no equivalent in the MDHs (here termed the tetramerisation interface). Although the LDH and MDH dimerisation interfaces are spatially equivalent they are made up of different amino acids and have different sets of interactions.

All of the LDHs and MDHs bind NAD as a cofactor in addition to either lactate ($COO^- \cdot CHOH \cdot CH_3$) or malate ($COO^- \cdot CHOH \cdot CH_2 \cdot COO^-$). As the main difference between the two substrates is malate's additional carboxyl group, one of the main specificity determinates of the enzymes is an additional positively charged arginine in the active site of MDH, the equivalent of which is an uncharged glutamine in the active site of LDH. LDH can be changed into MDH by mutation of this glutamine into arginine.[53] Apart from this, the binding sites both for substrate and cofactor are similar.

We aligned the sequences of the LDH/MDH family and partitioned the sequence into a phylogenetic tree using the neighbour-joining method as implemented in TraceSuiteII.[54]

When the entire, undivided, family is analysed, we find that the cofactor and substrate binding sites score highly. This is because these sites are common to all sequences. The exception is the specificity-determining Arg/Gln, which scores only 0.9. However, this residue has a high score when the sequences are broken down into substrate-specific sub-families: using only MDH sequences results in a score of 1.8 and only LDH sequences gives a score of 2.2.

Sequences were further subdivided into the three LDH and two MDH families described above, with bacterial sequences removed. The isolated protamer was used. These sub-groups contain only those sequences and structures specific to a particular sub-set of interface-forming residues. The result of this sub-division of sequences is that residues in the dimerisation interface additionally score highly with the top 5% of residues consisting only of residues in the binding and oligomerisation sites. The LDH tetramerisation interface also has high scoring residues, although results for the LDH families are more equivocal due to the lack of sequence divergence in the LDH sequence sub-families.

Thus, specificity-determining residues, both those that determine substrate specificity and those that determine oligomerisation interface specificity score highly only when considered in sub-groups which share common interface residues.

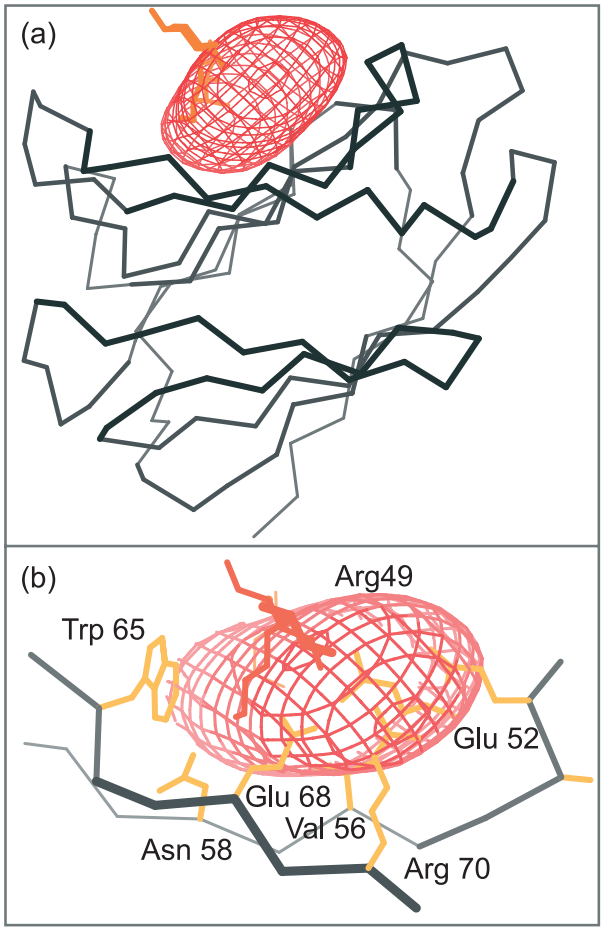

Figure 6. Contouring of the combined score for S-lectin (pdb code 1hlc) with sugar-binding residues and bound sugar shown.

Prediction of functional sites in proteins other than enzymes

Our method was first applied to enzyme active sites. However, since the method uses evolutionary information, it should be universally applicable to all other types of functional sites.

The sugar binding sites of lectins are able to bind a wide range of sugars and are rather flatter than enzyme active sites. Nevertheless, they too have residues that are more conserved than would expected from their amino acid type and environment. Figure 6 shows that the sugar-binding site of the S-lectin (pdb code 1hlc) is correctly identified.

Insulins and related hormones form a complex set of homo-oligomers. Human insulin has dimerisation and hexamerisation interfaces, and additionally complexes with its receptor. Within the insulin superfamily there are several sub-groups; for example the insulins/IGFs, bombyxin and the relaxins, each of which is specific for its own receptor and has distinct homo-dimerisation interfaces, although homo and hetero-complex interfaces overlap.[55,56]

A phylogenetic tree was built and the sequences aligned in each sub-branch. The sub-groups corresponded to insulin, insulin-like growth factor (IGF), bombyxin and relaxin. Table 6 shows the receptor-binding residues (red highlighted text), the homo-oligomerisation residues (black highlighted text) and the residues with the highest divergence scores (coloured background). Conservation scores are substantially similar.

Some interacting residues are conserved between different sub-groups, whereas others are sub-group specific. There is a good correlation between interface-forming residues and high-scoring residues. Particularly striking is relaxin, which has almost all of the high-scoring residues in the B chain, and all of the interface residues in the B chain. This is in contrast to the other sub-groups that have both high scoring and interface residues in both chains.

When compared to early predictions of the insulin receptor-binding region, an apparent false positive is seen at B6 Leu. However, the more recent observation that a human insulin mutation of Leu6 to Ala results in a 20-fold decrease in receptor binding affinity[57] argues that this residue should be considered functional.

A blind prediction

The strength of any prediction technique lies in its ability to make genuine predictions, which may be proved correct or incorrect with further knowledge. We have chosen, therefore, to predict the functional site of two related, hypothetical proteins. These predictions cannot currently be verified, but we wish them to be recorded in the literature so that the true accuracy of our methods may be judged in the future.

Hypothetical protein TM1083 (SwissProt id

Table 6. Insulin

A-chain	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Insulin				G	I	V		Q		C	T	S	I	C	S	L	Y	Q	L	E	N	Y	C	N
IGF-1				G	I	V		E	C	C	F		S	C	D	L	R	R	L	E	M		C	A
IGH-2				G	I	V		E	C	C	F		S	C	D	L	A	L	L	E	T		C	A
Bombyxin				G	I	V		E	C	C	L	R		C	S	V	D	V		L	S	Y	C	
Relaxin		L	Y	S	A	L	A	K	C	C	H	V	G	C	T	K	R	S	L	A	R	F	C	

B-chain	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Insulin				F	V	N	Q	H	L	C	G	S	H	L	V	E	A	L	Y	L	V	C	G	E	R	G	F	F	Y	T	P	K	A
IGF-1					G	P	E	T	L	C	G	A	E	L	V	D	A	L	Q	F	V	C	G	D	R	G	F	Y	F	S	R	P	A
IGF-2		A	Y	R	P	S	E	T	L	C	G	G	E	L	V	D	T	L	Q	F	V	C	G	D	R	G	F	Y	F	S	R	P	A
Bombyxin		Q	P	Q	A	V	H	T	Y	C	G	R	R	H	L	R	T	L	A	D	L	C	W	E	A	G	V	D					
Relaxin	S	W	M	E	E	V	I	K	L	C	G	R	E	L	V	R	A	Q	I	A	I	C	G	M	S	T	W						

Black text, dimerisation/hexamerisation interface. Red text, receptor binding. Green background, high scoring at family level. Orange background, high scoring at both family and subgroup level. Blue background, high scoring at subgroup level only.

Q9X0H1) is a conserved protein from the bacterium *Thermotoga maritima*. The crystal structure has been solved by the Joint Centre for Structural Genomics and found to have a new fold, consisting of two central α-helices flanked by two antiparallel β-sheets (pdb code 1j5u). The protein is a dimer, with the first β-strand undergoing strand exchange with the symmetry-related subunit. Hypothetical protein Mth1598 (SwissProt id O27635) is a conserved protein from the archaebacterium *Methanobacterium thermoautotrophicum*. The solution structure has been solved by Yee *et al.*[58](pdb code 1jw3). It is a monomer, structurally related to TM1083, but without the strand exchange. There is 20% sequence identity between the two proteins. Yee *et al.* suggest that there is structural homology with heat shock protein 33 (pdb code 1i7f) and ribosomal protein S8 (pdb code 1fka chain H) and tentatively assign a function of an RNA binding protein. They point out, however, that Mth1598 does not have the surface properties of an RNA binding protein. It should also be noted that, although the 1i7f, 1fkaH and 1jw3 have similar spatial arrangements of secondary structure, the connectivity differs in all three proteins and so they are unlikely to be related. Functional assignment on this basis should therefore be treated with some caution.

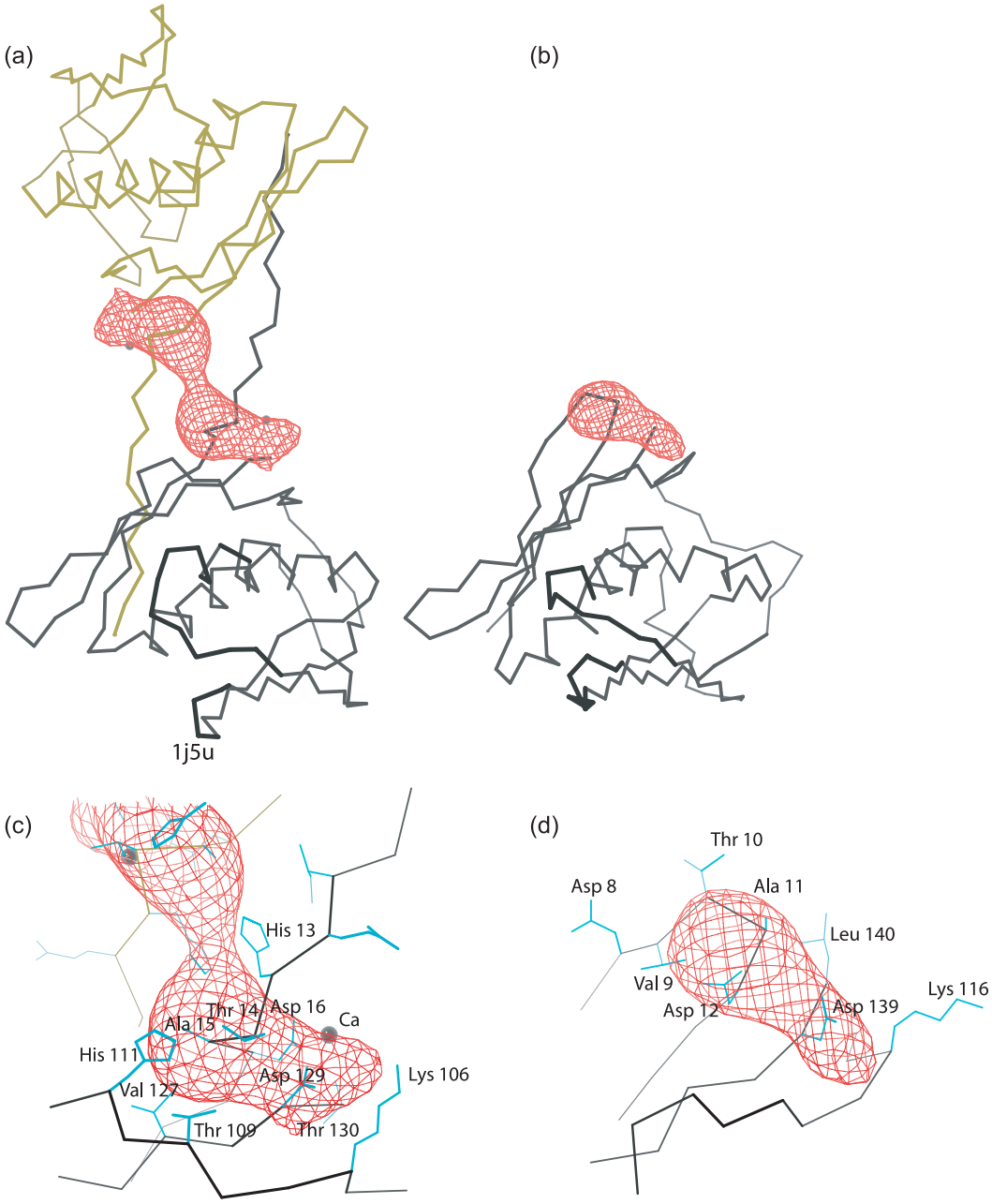

Figure 7. Blind prediction of the functional site of hypothetical proteins 1j5u and 1jw3 indicated by red contours. For 1j5u chain B is shown in yellow. The bottom panels show an enlarged view of the predicted site. Residues in the contour and neighbouring polar residues are shown.

All proteins with sequence similarity to TM1083 and Mth1598 (identified by PSI-BLAST) are hypothetical. The sequences fall into the PFAM[8] family DUF101, which contains only hypothetical proteins of unknown function.

We have calculated the conservation score for these two proteins and mapped it onto the structures. Functional sites are predicted in equivalent parts of the structures. For TM1083 (pdb code 1j5u) this site is between the two subunits and includes the residues that bind the Ca^{2+}. For Mth1598 this site is on a surface protrusion between β-strands 1 and 2 and β-strands 6 and 7. High-scoring residues include both conserved surface charges and hydrophobic residues. The residues in the region of the predicted functional site are shown in Figure 7.

Discussion

Evolutionary constraints on protein sequence can arise from a variety of sources, and can have differing strengths.[55] We have shown that analysis of these restraints can be used as to identify functional sites in proteins. Due to the ubiquitous nature of Darwinian evolution in determining protein sequence these techniques can be used to identify active sites, sugar binding sites and protein–protein binding sites. They can also be used to identify the binding sites of both homo- and hetero-protein complexes, in contrast to methods that use physical information. Only for a small number of proteins of the immune system, such as antibodies that have their binding sites produced by recombination of VDJ genes and somatic mutation, are these methods not applicable.

We have found that definition of protein functional sites is surprisingly difficult. Catalytic residues of enzymes are normally defined as those that directly take part in the reaction mechanism. These are widely reported in the literature, and we have been able to identify these for all of our test sets. Substrate binding residues are often understood and reported less well, and we have only been able to define these for the jack-knife set, due to the lack of protein–ligand complexes available and the difficulty of determining whether those that are available accurately represent the set of residues which bind the natural substrate. Residues that hold together an active site or stabilise the active conformation may also be thought of as functional, but are rarely reported as so in the literature. In principle, however, all of these residues can be identified by our scoring functions.

An even more generous definition of function residues is possible, for we must bear in mind that for most proteins to function they must attain and maintain the correct fold, be soluble and resist proteolysis. Residues that contribute to any of these roles may be thought of as functional residues. However, it appears that such evolutionary restraints do not in general conserve amino acids across families and they are unlikely to play a significant role in defining scores in the approach described here.

Our results differ with the three data sets. Most strikingly the number of false positives is substantially lower for the jack-knife set than for the other two. There are two possible explanations for this. It is possible that we were fortunate with our choice of the initial jack-knife set, which represents an unusual set of proteins. This would mean that our method genuinely works less well than the results from the jack-knife set suggest when more extensively tested. An alternative explanation is that the method works equally well on all sets, and only the information with which we benchmark our techniques varies. We think that this is most likely, given that the jack-knife set was chosen specifically because they are well characterised. Additionally, a wide range of protein–ligand complexes are available, from which we could choose that bind in the same mode as the substrate, and use this information to identify ligand-binding residues. Where we use the lower level of functional annotation for our jack-knife set, derived only from the literature and databases the proportion of false positives is similar to that of the other two datasets. It should be noted that the average number of residues identified as functional in the jack-knife set is 10.9, whereas the figure is 4.8 and 4.7 for the "new" and "non-independent" sets, respectively. For these reasons, we feel that the results for the jack-knife set are likely to be the most accurate.

Our techniques are based partly on identifying unusual evolutionary constraints on sequences, i.e. some degree of unexpected conservation. It has been shown that at the superfamily level of clustering, that there is a high degree of plasticity in the position and composition of enzyme active sites.[27,28] This is particularly the case when specificity changes within a family or superfamily, and even more so when different members of a superfamily catalyse different reactions, although these often have a common feature in substrate, intermediate or product. Our analysis shows that to identify correctly regions where specificity changes across a family or superfamily it is helpful to separate sequences into subfamilies, which can be done by phylogenetic analysis. This is in accordance with the results of others.[32–34,37,38] There is a compromise, however, between having only proteins of a given sub-class and therefore the same specificity and having proteins with enough sequence diversity to allow the conserved, evolutionarily restrained residues to be distinguished from those that have not had sufficient time to evolve since divergence.

Evolutionary restraints operate, at the organism level, on phenotype. At the molecular level, this translates to molecular function. Molecular function depends crucially on the three-dimensional arrangement of atoms. Thus, a protein must both have appropriate amino acid residues and position these appropriately in three dimensions. Thus,

analysis of both structural conservation, in addition to sequence conservation is valuable. We have found that the combination of a scoring system based on each of these factors gives the most accurate identification of protein functional sites.

Evolutionary constraints lead to conservation not only of sequence but also of structure.[29–31] For some proteins in our test set we are able to correctly identify the functional site on this basis. The combination of sequence and structure-based scoring systems seems to have an advantage of each alone.

The catalytic Asp102 of the serine proteinases is a buried charged residue. This is expected to be highly conserved for structural reasons.[42,43] Additional constraints from function do not lead to a significant increase in conservation of an already highly conserved residue and thus we are not able to identify any addition evolutionary constraints. Nevertheless, the advantages of the environment-specific substitution tables greatly outweigh the disadvantages and the three-dimensional smoothing function means that a change in a single residue is unlikely to greatly alter the position of the predicted functional site, although it may change the inclusion or otherwise of that residue in a contour of a specific level.

We must assume that the sequence alignment is accurate and the aligned residues structurally equivalent. We also assume that the environment of the amino acid does not change greatly. This is usually the case for residues in functional sites. We do not know, however, how true this is in the general case, nor are the evolutionary restraints on the type of environments used here very well understood. An understanding of the effects would greatly aid our understanding of evolutionary constraints on structure.

Methods

Database of aligned proteins

Environment-specific substitution tables were derived from the HOMSTRAD database[9]† using the SUBST program (K. Mizuguchi, unpublished results) as described.[42,43] At the time of compilation of the Tables, the database consisted of 3022 structures, grouped into 907 families and aligned on the basis of structure using MNYFIT[59,60] and COMPARER.[61] Non-environment-specific substitution tables were calculated by taking the mean of all environment-specific tables. The substitution tables were derived from 706 structures in 177 families.

The amino acid features used to define the amino acid environments were main- chain conformation, solvent accessibility and hydrogen-bonding class. Main-chain conformational classes were defined as α-helix, β-strand, coil and positive ϕ-angle (four classes); solvent accessibility was defined as buried (<7% of area accessible to a 1.4 Å probe) or accessible (two classes); hydrogen-bonding classes were defined according to the presence or absence of side-chain to main-chain carbonyl hydrogen bond, side-chain to main-chain amide hydrogen bond or side-chain to side-chain hydrogen bond ($2^3=8$ classes). In total this resulted in a total of $4\times2\times8=64$ classes.

By mapping HOMSTRAD database families and Enzyme structure database‡ families, 560 enzyme families were retrieved from HOMSTRAD. The following were not used: single member families, families that have their domains classified in different HOMSTRAD families and families that do not have enough information about their active sites in the literature or in SwissProt.[48] The final data set contains 244 families with well-defined active site residues.

The jack-knife sub-set of this data set was extracted. This consisted of ten families containing 136 structures and 1331 sequences of unknown structure. These families were removed while deriving the substitution tables. This set was selected because they have multiple high-resolution structures and at least one member of which had a well-defined structure of an inhibitor or substrate complex. The jack-knife set is detailed in Table 1.

A sub-set consisting of 392 structures (154 families out of which 106 are single- domain families and 48 are multi-domain families) was additionally selected to test the method. These structures (the "new" set) were selected such that they were added to HOMSTRAD after the substitution tables were calculated. None of them shares obvious sequence similarity (BLAST, *E*-value <0.1) to any of the structures that were used to derive the environment-specific substitution tables. Also used were 24,652 sequences of unknown structure.

Another sub-set consisting of 346 structures (80 families out of which 52 are single-domain families and 28 are multi-domain families) was selected to test the method. Also used were 10,912 sequences of unknown structure. These are the families that are related to or used to derive the substitution table and are included for completeness. This set is termed the "non-independent" set.

Definition of active site residues in the benchmark set

Residues were defined as being in the active site of an enzyme if they had been described as being catalytic or ligand binding in the literature, they were part of a PROSITE motif[44–46] or were listed in the ACT_SITE records of SwissProt. Additionally, residues from the jack-knife set were defined as functional if they contacted a ligand in the crystal structure. A contact was defined by the program PROBE[62] after addition of hydrogen atoms with REDUCE.[63] The default probe radius of 0.25 Å was used. In each case, the complex was checked to ensure the ligand bound in the same mode as the natural substrate. Because of the lack of suitable protein–ligand complexes and lack of information on natural binding modes, this information was not available for the new and non-independent data sets.

Collection and alignment of sequence homologues

In addition to proteins of known structure, other homologous sequences were used. These were collected, and aligned against the HOMSTRAD family, using PSI-BLAST[49] as implemented in the HOMBLAST utility (J. Shi, unpublished results). HOMBLAST aligns the homologous sequences against the HOMSTRAD

† http://www-cryst.bioc.cam.ac.uk/homstrad

‡ http://www.ebi.ac.uk/thornton-srv/databases/enzymes

structural alignment by taking each sequence in the alignment, running PSI-BLAST to collect homologues and combining them in a single file. From this alignment, clusters of sequences are made, where all the sequences within each cluster have greater than or equal to 80% identical sequences and none of the sequences between clusters are more than 80% identical. To avoid redundancy, only one protein sequence from each cluster was taken to build a multiple sequence alignment. This was achieved using FUGUESEQ (J. Shi, unpublished results).

Sequence conservation score

Two different methods for quantifying sequence conservation are used. With each scoring system, we attempt to compare the substitutions from the environment-specific substitution tables with the observed substitutions at a given position in the alignment. By quantifying the difference between these values we aim to identify extra restraints on evolution above those due the amino acid type and the environment.

The first scoring system, termed the conservation score is a modification of that of Rodionov & Blundell.[64] This score quantifies the degree of sequence conservation at an alignment position compared to the average conservation.

From the conservation analysis of residues in homologous sequences, the average conservation measure in a whole family at the tth position, $Con\_seq(t)$, is calculated:

$$Con\_seq(t) = \sum_{s=1}^{Sf} Con\_seq(t,s)/Sf$$

Here the average conservation measure $Con\_seq(t,s)$ is calculated for the tth position and measures the similarity of the residues of the sth sequence with the rest of the family sequences:

$$Con\_seq(t,s) = \sum_{r=1}^{Sf} Q(i(s),j(r))/Sf$$

All sequences in a family are aligned, where the number of such sequences is Sf. $Q(i(s),j(r))$ is the amino acid residue-similarity measure calculated for a pair of residues of ith and jth type. Types of residues i and j correspond to the types of residues at sth and rth aligned sequences at the tth position. $Q(i(s),j(r))$ is calculated according to:

$$Q(i,j) = \log(S_{\text{obs}}(i,j) + eps)/(S_{\text{exp}}(i,j) + eps)$$

Where eps is a positive small number (here $eps = 0.001$) which prevents taking the logarithm of zero when $S_{\text{obs}}(i,j)$ or $S_{\text{exp}}(i,j) = 0$.

$S_{\text{exp}}(i, j) = m_{ij}$ and is taken from the environment-specific substitution tables. This represents the substitution frequency seen on average for the exchange of amino acids i and j with this environment in the HOMSTRAD database of sequence alignments.

The observed substitutions between amino acids of type i and j, is given by $S_{\text{obs}}(i, j) = m_i P_j$ where m_i is the number of occurrences of amino acid i at this alignment position and P_j is the probability of finding amino acid j at this position P_j. The expected substitutions were estimated on the basis of the supposition that the probability to substitute the jth amino acid residue in a sequence by any other residue $P_j = m_j/m$ is random and uniform. It depends only on the ratio of the observed number of substitutions for amino acids of jth type m_j to the total number of substitutions, m, in a data set. At each alignment position the probability of amino acid distributions was calculated and termed the "observed substitution pattern".

The second sequence-based score, termed the "divergence score" quantifies the overall difference, or divergence, between the observed and predicted substitution probabilities. For the observed substitution pattern, the probability of a gap ($P_p(GR)$), at a position t, in the alignment was shared among the 20 amino acids (since any amino acid can occupy the gap region):

$$P_t(GR) = \sum_t GR/N$$

where N is the total number of homologous sequences.

The observed substitution pattern at position t for an amino acid x_i:

$$P_t(x_i) = \sum_t x_i/N$$

when $P_t(GR) = 0$:

$$\sum_t x_i/N + P_t(GR)/20$$

when $P_t(GR) \neq 0$

For each of the protein sequence of the HOMSTRAD structural alignment, the predicted substitution pattern of each of the 20 amino acids, at each position t was derived from the environment-specific substitution table, by taking its residue type and the environment in which it occurs. Taking the average over the number of structures available in the family, the predicted substitution pattern at each position for each of the 20 amino acids was calculated.

Given two distributions P and Q, the commonly used measure of statistical similarity between two arbitrary probability distributions, is the Kullback–Leibler (D^{KL},[65]) divergence defined as:

$$D^{\text{KL}}[P||Q] = \sum_i Px_i \log(Px_i/Qx_i) \text{ always} \geq 0$$

with equality if and only if:

$$P(x_i) = Q(x_i) \text{ for all } i = 1 \text{ to } 20$$

As this measure has the disadvantages of being asymmetric and unbounded, a better measure of statistical similarity as defined by Jensen–Shannon (D^{JS},[66] as suggested by Yona & Levitt.[67]) was used to find the divergence score between the two distributions P ("observed substitution pattern") and Q ("predicted substitution pattern").

Given two (empirical) probability distributions P and Q, for every $0 \leq \lambda \leq 1$, the λ-JS divergence is defined as:

$$D_\lambda^{\text{JS}}[P||Q] = \lambda D^{\text{KL}}[P||R] + (1-\lambda)D^{\text{KL}}[Q||R];$$

$$0 \leq D_\lambda^{\text{JS}}[P||Q] \leq 1$$

where:

$$R = \lambda P + (1-\lambda)Q$$

and:

$$\lambda = n/(n+m)$$

where n is number of structures available in the family, and m is total number of structures available in HOMSTRAD/total number of families in HOMSTRAD. R is the common source distribution of both distributions P and Q, with λ as a prior weight. This measure is

symmetric and ranges between 0 and 1, where the divergence for identical distributions is 0. The divergence score is calculated for each position t of the multiple sequence alignment.

If sequences of unknown structure are used, the average environment from the other proteins in the family was calculated and assigned to the sequences.

Sequences are given equal weighting whether or not they have experimentally determined structures or have their environments implied from the proteins of known structure.

Conservation of main-chain position

In order to calculate the conservation of main-chain conformation it was necessary to superimpose the proteins. All proteins in the family were superimposed onto the protein of interest. Residues are considered structurally equivalent if they are aligned in the input alignment file. The 5% of residues with highest conservation score were chosen as the initial equivalence set with superposition based on only the C^{α} atoms. The equivalence set is expanded by iteratively adding the residue pair with shortest pair-wise C^{α}–C^{α} distance. The equivalence set is expanded until the RMSD value is greater than 1.5 Å or all residues are included. This procedure may be thought of as the inverse of the sieving algorithm of Irving *et al*.[29] All protein structures included must be in the same liganded state (i. e. all liganded or all unliganded).

Based on this superposition, the structural conservation score *Con_struc* was calculated at the tth position, *Cons_strc*(t) is calculated:

$$Cons\_struc(t) = \sum_{s=1}^{Sf} Cons\_struc(t,s)/Sf$$

Here the average conservation score *Cons_struc*(t,s) is calculated for the tth position and measures the conservation of the sth structure with the rest of the family structures:

$$Cons\_struc(t,s) = 1/\log(dis + eps)$$

Where *dis* is the distance in Ångstrom units between equivalent C^{α} atoms.

Both sequence and structure-based scores were converted to Z-scores using the formula:

$$Z = (\text{score} - \text{mean})/\text{standard deviation}$$

Combined scores

Z-scores were combined using empirically determined weights of 0.65×sequence-based score+0.35×structure-based score.

Contouring of scores

A representative protein was chosen from the family. The score associated with each residue position was assigned the three-dimensional coordinate of the atom most likely to be involved in the function of the side-chain. These are detailed in Table 4. Where more than one atom is listed the mean coordinate position is used. The values were smoothed and contoured using Kin3Dcont†.

This places a three-dimensional Gaussian mask at the position of the chosen coordinate. The mask was chosen such that it had a standard deviation of 3 Å, an integral equal to the combined score, and was sampled at grid points spaced by 1 Å. At each grid point, the score is summed: thus residue scores sum with those neighbouring scores if they are close in three dimensions (within the expanse of the mask). The summed mask scores are contoured and the contours viewed with the vector-viewing program MAGE.[68,69]

Automatic prediction of functional residues

The three-dimensional grid of points was calculated with kin3Dcont as described above. A maximum Z-score was chosen such that the number of grid points above this Z-score is greater than or equal 400. Then the cut-off is set:

$$\text{cut} - \text{off} = \text{mean} - (\text{maximum } Z - \text{score}) * \sigma$$

where σ is the standard deviation.

All grid points with Z-score above the cut-off were clustered, and the clusters ranked in size. Clusters separated by less than 3 Å were merged. A residue is predicted to be functional if it has an atom within 0.8 Å of a grid point in the largest cluster. These assignments were used to determine correctly predicted functional residues ("true positives"), missed functional residues ("false negatives") and residues predicted to be functional which were not ("false positives").

† Word, J. M. (2000) Department of Biochemistry, Duke University, Durham, NC.

Acknowledgements

We thank Ross Munro for providing information about characterised enzyme active sites and Kenji Mizuguchi for useful suggestions. S.C.L. was supported as a Wellcome Trust Fellow of Mathematical Biology. V.C. is supported by the Nehru Cambridge Trust and Overseas Research Studentship. L.C. was supported by the Cambridge Overseas Trust and an Overseas Research Studentship award.

References

1. Blundell, T. L., Sibanda, B. L., Sternberg, M. J. & Thornton, J. M. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, **326**, 347–352.
2. Sali, A. (1998). 100,000 protein structures for the biologist. *Nature Struct. Biol.* **5**, 1029–1032.
3. Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T. *et al*. (1999). Structural genomics: beyond the Human Genome Project. *Nature Genet.* **23**, 151–157.
4. van Helden, J., Naim, A., Mancuso, R., Eldridge, M., Wernisch, L., Gilbert, D. & Wodak, S. J. (2000). Representing and analysing molecular and cellular function using the computer. *Biol. Chem.* **381**, 921–935.
5. Andrade, M. A. & Valencia, A. (1998). Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, **14**, 600–607.

6. Andrade, M. A., Brown, N. P., Leroy, C., Hoersch, S., de Daruvar, A., Reich, C. *et al*. (1999). Automated genome sequence analysis and annotation. *Bioinformatics*, **15**, 391–412.
7. Hofmann, K., Bucher, P., Falquet, L. & Bairoch, A. (1999). The PROSITE database, its status in 1999. *Nucl. Acids Res.* **27**, 215–219.
8. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R. *et al*. (2002). The Pfam Protein Families Database. *Nucl. Acids Res.* **30**, 276–280.
9. 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.
10. Kasuya, A. & Thornton, J. M. (1999). Three-dimensional structure analysis of PROSITE patterns. *J. Mol. Biol.* **286**, 1673–1691.
11. Fetrow, J. S., Godzik, A. & Skolnick, J. (1998). Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* **282**, 703–711.
12. Fetrow, J. S. & Skolnick, J. (1998). Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T_1 ribonucleases. *J. Mol. Biol.* **281**, 949–968.
13. Zhang, B., Rychlewski, L., Pawlowski, K., Fetrow, J. S., Skolnick, J. & Godzik, A. (1999). From fold predictions to function predictions: automation of functional site conservation analysis for functional genome predictions. *Protein Sci.* **8**, 1104–1115.
14. Fetrow, J. S., Siew, N., Di Gennaro, J. A., Martinez-Yamout, M., Dyson, H. J. & Skolnick, J. (2001). Genomic-scale comparison of sequence- and structure-based methods of functional prediction: does structure provide additional insight? *Protein Sci.* **10**, 1005–1014.
15. Di Gennaro, J., Siew, N., Hoffman, B. T., Zhang, L., Skolnick, J., Neilson, L. I. & Fetrow, J. S. (2001). Enhanced functional annotation of protein sequences *via* the use of structural descriptors. *J. Struct. Biol.* **134**, 232–245.
16. Herzberg, O. & Moult, J. (1991). Analysis of the steric strain in the polypeptide backbone of protein molecules. *Proteins: Struct. Funct. Genet.* **11**, 223–229.
17. Heringa, J. & Argos, P. (1999). Strain in protein structures as viewed through non-rotameric side chains: II. Effects upon ligand binding. *Proteins: Struct. Funct. Genet.* **37**, 44–55.
18. Elcock, A. H. (2001). Prediction of functionally important residues based solely on the computed energetics of protein structrue. *J. Mol. Biol.* **213**, 885–896.
19. Ota, M., Kinoshita, K. & Nishikawa, K. (2003). Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile and sequence conservation. *J. Mol. Biol.* **327**, 1053–1064.
20. Laskowski, R., Luscombe, N. M., Swindells, M. B. & Thornton, J. M. (1996). Protein clefts in molecular recognition and function. *Protein Sci.* **5**, 2438–2452.
21. Jones, S. & Thornton, J. M. (1997). Analysis of protein–protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121–132.
22. Honig, B. & Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science*, **268**, 1144–1149.
23. Zhu, Z.-y. & Karlin, S. (1996). Cluster of charged residues in protein three-dimensional structures. *Proc. Natl Acad. Sci.* **93**, 8350–8355.
24. Aloy, P., Querol, E., Aviles, F. X. & Sternberg, M. J. E. (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.
25. Stawiski, E. W., Gregoret, L. M. & Mandel-Gutfreund, Y. (2003). Annotating nucleic acid-binding function based on protein structure. *J. Mol. Biol.* **326**, 1065–1079.
26. Zvelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195**, 957–961.
27. Todd, A. E., Orengo, C. A. & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143.
28. Todd, A. E., Orengo, C. A. & Thornton, J. M. (2002). Plasticity of enzyme active sites. *Trends Biochem Sci.* **27**, 419–426.
29. Irving, J. A., Whisstock, J. C. & Lesk, A. M. (2001). Protein structural alignments and functional genomics. *Proteins: Struct. Funct. Genet*, **42**, 378–382.
30. Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO*, **5**, 823–826.
31. McPhalen, C. A., Vincent, M. G., Picot, D., Jansonius, J. N., Lesk, A. M. & Chothia, C. (1992). Domain closure in mitochondrial aspartate aminotransferase. *J. Mol. Biol.* **227**, 197–223.
32. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). The evolutionary trace method defines the binding surfaces common to a protein family. *J. Mol. Biol.* **257**, 342–358.
33. Lichtarge, O., Yamamoto, K. R. & Cohen, F. E. (1997). Identification of functional surfaces of the zinc binding domains of intracellular receptors. *J. Mol. Biol.* **274**, 325–337.
34. Madabushi, S., Yao, H., Marsh, M., Kristensen, D., Philippi, A., Sowa, M. E. & Lichtarge, O. (2002). Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **316**, 139–154.
35. Yao, H., Kristensen, D. M., Mihalek, I., Sowa, M. E., Shaw, C., Kimmel, M. *et al*. (2003). An accurate, sensitive and scalable method to identify functional sites in protein structures. *J. Mol. Biol.* **326**, 255–261.
36. Lichtarge, O. & Sowa, M. A. (2002). Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.* **12**, 12–27.
37. Armon, A., Graur, A. & 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.
38. Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18**, S71–S77.
39. 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.
40. Bork, P. & Koonin, E. V. (1998). Predicting functions from protein sequences—where are the bottlenecks? *Nature Genet.* **18**, 313–318.

41. Karp, P. (1998). What we do not know about sequence analysis and sequence databases. *Bioinformatics*, **14**, 753–754.
42. Overington, J., Donnelly, D., Johnson, M. S., Sali, A. & Blundell, T. L. (1992). Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.* **1**, 216–226.
43. Overington, J., Johnson, M. S., Sali, A. & Blundell, T. L. (1990). Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. R. Soc. Lond. B: Biol. Sci.* **241**, 132–145.
44. Burcher, P. & Bairoch, A. (1994). A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 53–61.
45. Bairoch, A. & Burcher, P. (1994). PROSITE: recent developments. *Nucl. Acids Res.* **22**, 3626–3627.
46. Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K. & Bairoch, A. (2002). The PROSITE database, its status in 2002. *Nucl. Acids Res.* **30**, 235–238.
47. 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.
48. Bairoch, A. & Apweiler, R. (1997). The SWISS-PROT protein sequence database: its relevance to human molecular medical research. *J. Mol. Med.* **75**, 312–316.
49. Altschul, S. F. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database programs. *Nucl. Acid Res.* **25**, 3389–3402.
50. Pearl, L. H. & Blundell, T. L. (1984). The active site of aspartic proteinases. *FEBS Letters*, **174**, 96–101.
51. Chapman, A. D. M., Cortes, A., Dafforn, T. R., Clarke, A. R. & Brady, R. L. (1999). Structural basis of substrate specificity in malate dehydrogenases: crystal structure of a ternary complex of procine cytoplasmic malate dehydrogenase, alpha-ketomalonate and tetrahydroNAD. *J. Mol. Biol.* **285**, 703–712.
52. Crawford, D. L., Constantino, H. R. & Powers, D. A. (1989). Lactate dehydrogenase-B cDNA from the Teleost *Fundulus heteroclitus*: evolutionary implications. *Mol. Biol. Evol.* **6**, 369–383.
53. Wilks, H. M., Hart, K. W., Feeney, R., Dunn, C. R., Muirhead, H., Chia, W. N. *et al*. (1988). A specific, highly active malate dehydrogenase by redesign of a lactate dehydrogenase framework. *Science*, **242**, 1541–1544.
54. Innis, C. A., Shi, J. & Blundell, T. L. (2000). Evolutionary trace analysis of TGF-beta and related growth factors: implications for site-directed mutagenesis. *Protein Eng.* **13**, 839–847.
55. Blundell, T. L. & Wood, S. P. (1975). Is the evolution of insulin Darwinian or due to selectively neutral mutation? *Nature*, **257**, 197–203.
56. Conlon, J. M. (2001). Evolution of the insulin molecule: insights into structure-activity and phylogenetic relationships. *Peptide*, **22**, 1183–1193.
57. Kristensen, C., Kjeldsen, T., Wiberg, F. C., Schaffer, L., Hach, M., Havelund, S. *et al*. (1997). Alanine scanning mutagenesis of insulin. *J. Biol. Chem.* **272**, 12978–12983.
58. Yee, A., Chang, X., Pineda-Lucena, A., Wu, B., Semesi, A., Le, B. T. *et al*. (2002). An NMR approach to structural genomics. *Proc. Natl Acad. Sci.* **99**, 1825–1830.
59. Sutcliffe, M. J., Haneef, I., Carney, D. & Blundell, T. L. (1987). Knowledge based modelling of homologous proteins. Part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng.* **1**, 377–384.
60. Sutcliffe, M. J., Hayes, F. R. F. & Blundell, T. L. (1987). Knowledge based modelling of homologous proteins. Part II: rules for the conformations of substituted sidechains. *Protein Eng.* **1**, 385–392.
61. Sali, A. & Blundell, T. L. (1990). Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.* **212**, 403–428.
62. Word, J. M., Lovell, S. C., LaBean, T. H., Taylor, H. C., Zalis, M. E., Presley, B. K. *et al*. (1999). Visualizing and quantifying molecular goodness of fit: small-probe contact dots with explicit hydrogen atoms. *J. Mol. Biol.* **285**, 1711–1733.
63. Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (1999). Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **285**, 1735–1747.
64. Rodionov, M. A. & Blundell, T. L. (1998). Sequence and structure conservation in a protein core. *Proteins: Struct. Funct. Genet.* **33**, 358–366.
65. Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York.
66. Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory*, **37**, 145–151.
67. Yona, G. & Levitt, M. (2002). Within the twighlight zone: a sensitive profile–profile comparison tool based on information theory. *J. Mol. Biol.* **315**, 1257–1275.
68. Richardson, D. C. & Richardson, J. S. (1992). The kinemage: a tool for scientific illustration. *Protein Sci.* **1**, 3–9.
69. Richardson, D. C. & Richardson, J. S. (2001). In *International Tables for Crystallography* (Rossmann, M. G., Arnold, E., eds), pp. 727–730, Kluwer Publishers, Dordrecht. Chapter 25.2.8.
70. Mirny, L. A. & Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **291**, 177–196.

Edited by J. Thornton

(Received 27 November 2003; received in revised form 20 July 2004; accepted 9 August 2004)

Note added in proof: Since submission of this paper it has been brought to our attention that Mirny and Shakhnovich[70] have presented an alternative approach to distinguishing structural and evolutionary constraints.