doi:10.1006/jmbi.2001.4870 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2001) **311**, 395–408

JMB

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

Patrick Aloy[1,2], Enrique Querol[1], Francesc X. Aviles[1] and Michael J. E. Sternberg[2]*

[1]*Institut de Biologia Fonamental and Departament de Bioquimica, Universitat Autonoma de Barcelona, Bellaterra 08193, Barcelona Spain*

[2]*Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, London WC2A 3PX, UK*

**Corresponding author*

A major problem in genome annotation is whether it is valid to transfer the function from a characterised protein to a homologue of unknown activity. Here, we show that one can employ a strategy that uses a structure-based prediction of protein functional sites to assess the reliability of functional inheritance. We have automated and benchmarked a method based on the evolutionary trace approach. Using a multiple sequence alignment, we identified invariant polar residues, which were then mapped onto the protein structure. Spatial clusters of these invariant residues formed the predicted functional site. For 68 of 86 proteins examined, the method yielded information about the observed functional site. This algorithm for functional site prediction was then used to assess the validity of transferring the function between homologues. This procedure was tested on 18 pairs of homologous proteins with unrelated function and 70 pairs of proteins with related function, and was shown to be 94% accurate. This automated method could be linked to schemes for genome annotation. Finally, we examined the use of functional site prediction in protein-protein and protein-DNA docking. The use of predicted functional sites was shown to filter putative docked complexes with a discrimination similar to that obtained by manually including biological information about active sites or DNA-binding residues.

Keywords: Bioinformatics; structural genomics; protein function prediction; active sites

Introduction

A major strategy in the assignment of biological activity to proteins in sequenced genomes is the transfer of function between sequence homologues.[1,2] However, homologues, particularly when the sequence identity is below 25%, need not have related activities[3–6] and, consequently, functional inheritance leads to only tentative annotations.[7,8] Here, we introduce a strategy that exploits knowledge of the three-dimensional structure of the protein superfamily to establish whether functional inheritance within the superfamily is likely to be valid.

Several strategies are used to annotate the protein products in genomes, e.g. see reviews by Bork *et al*.[1] and by Bork & Koonin.[2] In the absence of direct experimental information, generally one searches for sequence matches to databases of patterns or motifs of characterised functions such as PFAM,[9] SMART,[10] PROSITE[11] or BLOCKS.[12] Without a match to such pattern/motif databases, simply finding a medium or weak sequence homology to a protein of known function is not sufficient for a confident annotation.[7,8]

Recently, the problem of functional inheritance has been evaluated using the structural classifi-

Present addresses: P. Aloy, Biocomputing, EMBL, Meyerhofstrasse 1, Heidelberg D-69117, Germany; M. J. E. Sternberg, Department of Biology & Biochemistry, Imperial College of Science, Technology and Medicine, London SW7 2AY, England.

Abbreviations used: PDB, Protein Data Bank; SCOP, structural classification of proteins; EC, enzyme commission.

E-mail address of the corresponding author: m.sternberg@ic.ac.uk

cation of proteins (SCOP) database,[13] which exploits commonality of structure to group remote homologues into a single superfamily. Russell *et al*.[3] showed that 10 % of remote homologues (undetectable by BLAST) in a SCOP superfamily have quite different functions. Hegyi & Gerstein[4] examined proteins in SCOP. They identified a set of homologous proteins with different functions: either two different enzyme activities, as formalised by differences in the last three digits of the enzyme classification (EC) number,[14] or one enzyme and one non-enzyme. Subsequently, Wilson *et al*.[5] and Devos & Valencia[6] quantified the relationship between the degree of similarity in sequence between two proteins and the likelihood that they have a common function. Below 25 % identity, 25 % of pairs of homologous proteins can have quite different functions.

One guide to the commonality of activity between homologues is the evolutionary conservation of functionally important residues. Furthermore, sequence conservation within a subfamily can suggest residues important for specificity.[15–17] The availability of the three-dimensional structure for one member of the family can assist in the prediction of functionally important residues. Lichtarge *et al*.[18,19] have introduced the evolutionary trace approach, in which one searches visually for spatial clusters of evolutionarily conserved residues to suggest probable functional sites. They originally demonstrated that the evolutionary approach is successful on a few systems. Subsequently they applied the method to identify functional surfaces of intracellular receptors.[20] More recently, the strategy has been extended to consider correlated mutations.[21,22]

Here, our objective was to evaluate the reliability of using the conservation of a functional site, predicted by the evolutionary trace method, as an indication of the validity of functional inheritance. The first step was to develop a fully automated method to perform the evolutionary trace method.[18,19] The method was then benchmarked in terms of required sequence divergence, and the resultant selectivity and specificity of the prediction. The optimisation and benchmarking of this approach is increasingly important. The advent of structural genomics projects (e.g. see Burley *et al*.[23]) will lead to the determination of the conformation of numerous proteins prior to knowledge of their function. Identification of probable functional regions can focus subsequent experimental work such as mutagenesis. Once the evolutionary trace algorithm was benchmarked, we applied it to the problem of functional inheritance. Finally, we evaluated the use of the prediction of the location of functional sites to assist in filtering putative complexes generated by our macromolecular docking algorithm, which starts with coordinates of the unbound components.[24–27] This docking study considered both protein-protein and protein-DNA complexes.

Benchmarking the Prediction of Functional Sites

Algorithm

Define the observed functional site

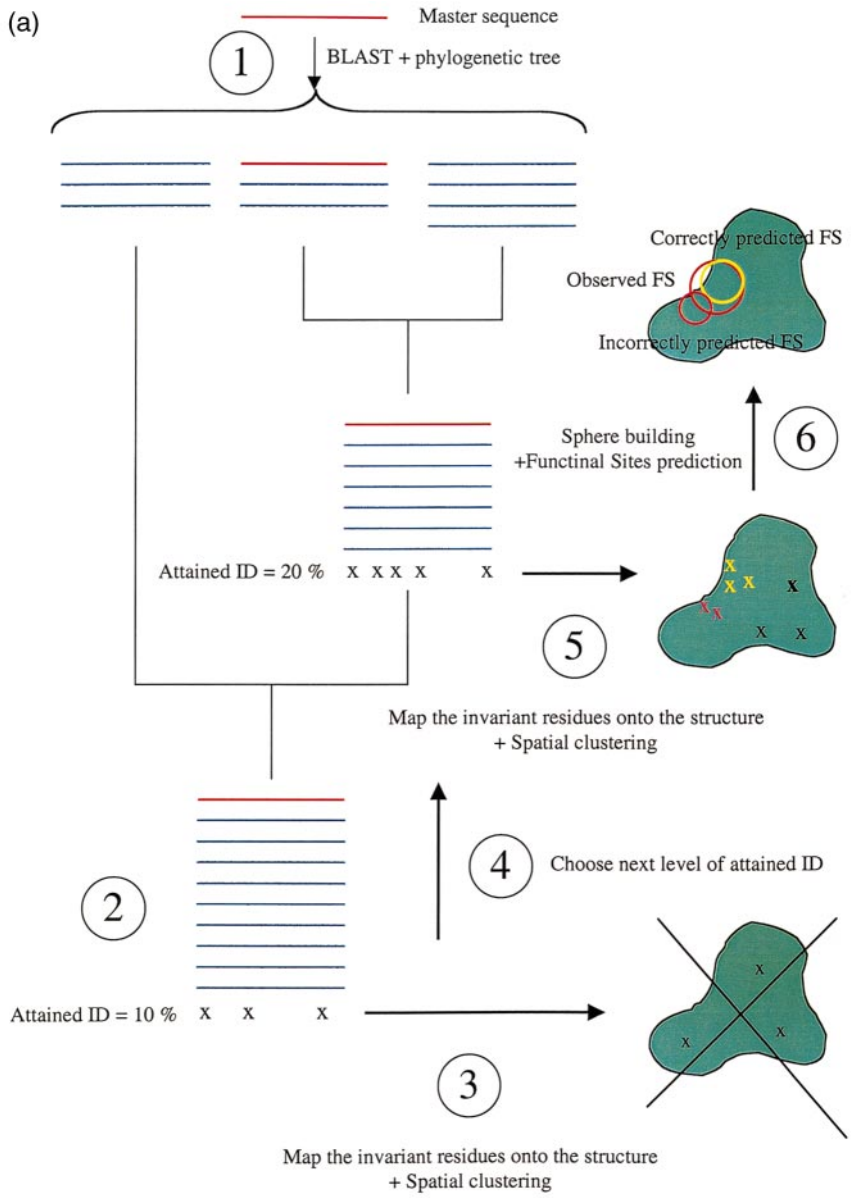
The aim is to implement and to benchmark an automatic clustering method to predict functional sites in proteins (see Figure 1(a)). In this study, we consider that the functional site of a protein is a region that performs any of a diverse set of activities, including acting as an enzyme active site or being a binding region for a small molecule or a macromolecule. The objective is to establish whether a single computational strategy can be used for a diverse set of activities rather than developing different approaches for different classes of proteins. The information about the residues forming these functional sites was extracted from the ACTSITE and SITE records as identified in the RSCB protein data bank (PDB).[28] The information in these records is of variable quality but has been reported by the group that determined the coordinates and thereby is the result of an in-depth, structure-based knowledge of activity. We consider that this approach, whilst not ideal, provides an automated and unbiased strategy required for a benchmark on a large dataset.

An initial set of ~1800 proteins was obtained by taking all the protein chains with a description of functional residues from the March 1999 release of the PDB.[28] The size of our dataset is in accord with the 2234 proteins that have an EC annotation that were used in another recent study[6] (D. Devos, personal communication). To build up a non-redundant subset of proteins, the program OBSTRUCT[29] was used on the initial set of (1800 proteins. The program identifies one representative for all homologous protein chains with more than 25 % sequence identity with one another. OBSTRUCT selects the homologue with the best crystal resolution as the representative chain. The resultant set consisted of 106 chains with less than 25 % mutual sequence identity (for details, see Materials). The size of our resultant database of 106 chains is in accord with the level of reduction achieved by PDB_SELECT at 25 % identity.[30] The proteins analysed were 95 enzymes and 11 non-enzymes.

We define the observed functional site as a single sphere that encloses the residues identified from the PDB as involved in the activity of the protein. The geometric centroid of the residues extracted from the PDB is used as the centre and the largest distance from the centre to any C^{β} atom (C^{α} for glycine) of the residues is used as the radius.

Find homologous sequences

The sequence of the master protein of known structure is used as a query in a sequence-



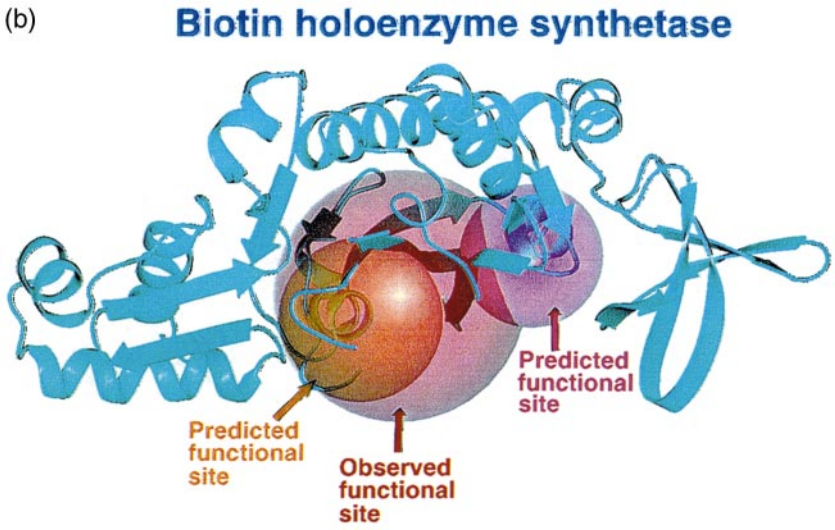


Figure 1. Overview of the method and evaluation of the results. (a) A schematic of the mapping of conserved sequence residues onto a structure. The attained identity in the alignment is shown. In this example, the lowest level of the attained identity in the multiple alignment does not yield a spatial cluster and consequently the next highest level is examined and, in the illustration, used to generate the cluster. (b) A typical result showing the predicted functional sites in biotin holoenzyme synthetase (1bia; EC 6.3.4.15). In red is the sphere corresponding to the observed functional site, which contains the 11 residues described as catalytic. Note that this observed functional site is larger than the average. In yellow and magenta are the two predicted functional sites. The yellow functional site has an overlap of 100 %, it is completely inside the observed functional site sphere. The magenta one, despite not being considered a good prediction (32 % overlap), still identifies some true residues of the functional site. In this protein, there were 18 residues in the two predicted functional sites and of these, ten corresponded to the 11 observed functional residues.

homology search against the non-redundant (nrdb95) sequence database.[31] The database search method WU-BLAST[32] is used to collect sequence matches with *e*-values lower than 10^{-3} and a matching length greater than 50 % of the query sequence. This *e*-value cut off (without the additional length constraint) is expected to generate, on average, 0.01 erroneous match per query.[33] With just over 100 proteins being examined, there should only be one false homology included in the entire analysis. Inspection of all the multiple alignments did not reveal any obvious false homology.

Generate a multiple sequence alignment and a phylogenetic tree

A multiple sequence alignment is then generated with CLUSTALW[34] using the default settings (including the BLOSUM62 sequence similarity matrix). For technical reasons, the dendrogram in CLUSTALW was not used, but instead a nearest-neighbour hierarchical clustering algorithm[35] was implemented using sequence identity alone. Very similar results would be expected if CLUSTALW were run using the identity matrix to build up the phylogenetic tree. Once the dendrogram is obtained, the query sequence is traced along the tree and every time that a new sequence, or group of sequences, is added to its own cluster, a consensus sequence is constructed. At each alignment position, residues that are totally conserved (permitting one residue sequence error) are taken as invariant.

In this study the term attained identity is defined as the number of invariant residues in the multiple alignment divided by the length of the query, expressed as a percentage. Each node in the phylogenetic tree is associated with an attained identity for the multiple alignment (Figure 1(a)).

Predict the functional site

The invariant polar residues (D,E,H,K,N,Q,R,S,T) are mapped onto the master known structure. Then, spatial clustering[35] is applied to identify those invariant residues that are close in space. First, residues are clustered by single linkage, so a residue is added to the cluster if it is less than 8 Å from at least one other residue in the cluster. Then, for residues clustered by single linkage, one starts with the closest pair of residues in the cluster and the centroid is evaluated. The next closest residue to the centroid is added to the cluster provided the resultant centroid shifts by less than 7 Å from its previous location. This is repeated until no more residues are added to the cluster. Finally, a sphere containing all the clustered residues is built using the geometric centroid of the cluster as the centre and the largest distance from the centre to any C^{β} atom (C^{α} for glycine) of the clustered residues as the radius. This sphere and all the residues inside form the predicted functional site. Note that the predicted functional site includes all residue types within the sphere and not just the polar residues used for clustering.

If no cluster can be generated at the lowest attained identity (from the multiple alignment), the sequence dendogram is examined and the subset of invariant residues corresponding to the previous branch (at a higher identity) in the phylogenetic tree is used for spatial clustering (see Figure 1(a)). This can be repeated, moving up the tree until a cluster is obtained. Thus, there always is at least one spatial cluster of invariant polar residues identified by the approach.

The above methodology, in particular the choice of distance cut-offs, was developed by inspection of the catalytic triad of the serine proteases in the trypsin family. The cut-offs were not optimised on any of the subsequent data used in this study. For details of the dependency of the results on the use of C^{β} atoms and choice of the distance cut-offs, see Materials.

Functional sites prediction

For each protein, we evaluated the volume overlap between the sphere forming the observed functional site with each predicted functional site sphere. If the overlap was $\geqslant$50 % of the volume of the predicted functional site, then the predicted site was considered correct. The results described below are not altered substantially if this cut-off is changed to either 40 % or 60 % overlap (for more information, see Figure A1 on the associated web page†). Figure 1(b) illustrates one protein that yielded a correct functional site prediction.

Table 1 shows the accuracy of the functional site predictions for the 106 proteins. The results are banded by the attained sequence identity used to generate the prediction. We observe that for attained sequence identities >30 %, the method was of very limited use. This occurs for about 20 % of the protein chains considered (20 out of 106). In contrast, for the (30 % band, the percentage accuracy of correctly predicted functional sites is 51 %, with the average number of functional sites predicted per protein being 2.7. In the remaining analysis in this section, we consider only the results from the $\leqslant$30 % band. Clearly, a predicted site that has no overlap with the observed site is incorrect but this occurs for only about 25 % of the predicted sites. For the rest, although not being considered as correct, the prediction still yields some useful information about the location of the functional residues in the protein. Details of the results banded in 10 % sequence identity segments are available *via* the accompanying web page (see Figure A2†). These data show that there is a trend that, as the attained identity increases there is a progressive increase in the number of predicted functional sites

† Available online at http://luz.uab.es/biocomputing/patrick/functsites.html

Table 1. Functional sites prediction results

Attained %ID	No. of proteins	No. of predicted functional sites	% Predicted FS with > 50 % overlap	% Predicted FS with 0 % overlap
>30	20	376	14	58
⩽30	86	230	51	26

The first column gives the range of identity (%ID) attained in the multiple alignment, the second column the number of proteins, and the third column the total number of predicted functional sites (FS) at each identity level. The fourth column shows the percentage of correctly predicted functional sites (with an overlap between the predicted and the observed functional sites ⩾50 % by volume) and the fifth column shows the percentage of completely wrong functional site predictions (overlap = 0 %).

and a concomitant decrease in the accuracy of prediction.

The results for the protein chains examined are available on the web page (see Table A1). Figure 2 summarises the results for the 86 proteins in the ⩽30 % identity band. We consider that the method has found a functional site if at least one of the predicted functional sites is correct. For 79 % of the proteins, the method finds the correct location of the functional site. For 6 % of the proteins, the predictions are completely wrong. For the other 15 % of the proteins, at least one of the predicted sites still has some (but less than 50 %) overlap with the observed site and therefore the prediction identifies some of the residues involved in the function of the protein. Of the 86 proteins, 80 were enzymes, and for 62 of them, the correct location of the functional site was predicted. For all of the six non-enzymes, the functional site was identified.

These results are not due to predicting either a single large functional site or many small functional sites, which results in trivial overlap with the observed functional site (see Figure A3†). On the contrary, the volume of the predicted functional sites generally is smaller than the observed ones. On average, the volume of the observed functional site is 4 % of the volume of the protein (the volume of the protein is based on the sphere that encloses all C^{β} atoms). In contrast, the sum of the volumes of all the predicted functional sites per protein is, on average, only 4 % of the total volume of the protein. Only in three proteins was the sum of the predicted functional sites more than 10 % of the total volume of the protein. Inspection of the results for each predicted functional site (see additional Table A1 on the web page) shows that certain situations that, in principle, could occur are not observed. In particular, there is no occurrence where a predicted functional site is scored as a hit but none of the observed functional residues is actually enclosed by the predicted sphere. In addition, there is no case where the volume of the predicted site is more than twice that of the observed site and consequently the predicted site could never be scored as correct.

In order to assess the statistical significance of these results, we can estimate the probability (p_r) that a randomly chosen predicted functional site will overlap ⩾50 % by volume with the observed functional site. The volume of a predicted functional site is, in general, smaller than the volume of the observed functional site (Figure A2 on the web page). An upper estimate of p_r is that the centre of the predicted site is within the sphere of the observed site. Thus the upper bound on p_r is simply the fraction of the volume of the observed functional site compared to the total volume of the protein. The average of this fraction over all the proteins is 4 % compared to the observed accuracy of correctly predicted functional sites of 51 % (p_o) (Table 1). The significance of the difference between p_o and p_r is estimated from:

$$\frac{P_o - p_r}{\sqrt{p_r(1-p_r)/n}}$$

where n is the number of clusters. The value follows a normal distribution with mean 0 and

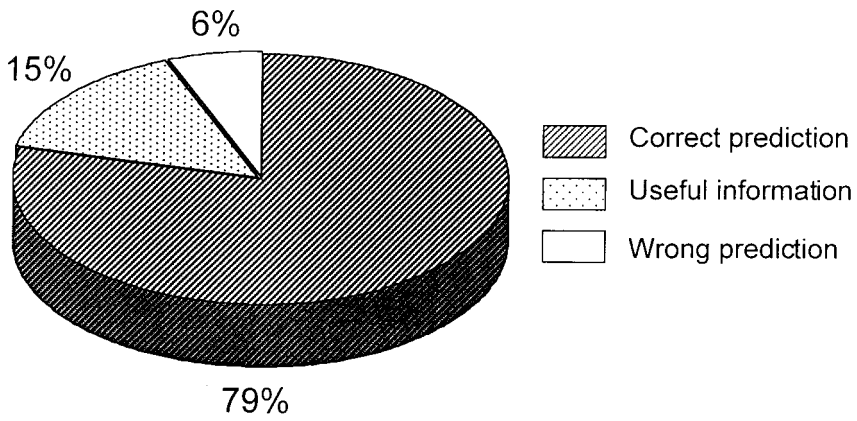


Figure 2. Pie chart showing the per protein accuracy in the ⩽30 % attained identity band.

standard deviation 1. The difference is significant at better than 10^{-5}. One could argue that because of stereochemical constraints, the predicted and observed spheres would occupy only part of the volume of the protein. Visual inspection of our results suggests that this is not a major problem. However, a worst-case scenario would be that the centroid is restricted to only one-third of the volume of the protein, then p_r becomes 12 % and still the difference is significant at better than 10^{-5}. We now estimate the probability (q_r) that by chance the observed site is identified correctly. We use an average of three predicted functional sites per protein and consider the worst case that these three sites do not overlap, In addition, we restrict the centroids to only one-third of the total protein volume. This leads to a q_r of 36 %. The observed accuracy (q_o) for predicting the observed sites is 79 % for 86 proteins. Again, the difference is significant at better than 10^{-5}.

We now consider how many of the residues defined as functional in the header of the protein coordinates entry are located inside the predicted functional site spheres. The standard measures of sensitivity and specificity are used:

$$\text{Residue sensitivity} = \frac{t_p}{t_p + f_n}$$

$$\text{Residue specificity} = \frac{t_p}{t_p + f_n}$$

where t_p = true positives, f_n = false negatives (under prediction), f_p = false positives (over prediction).

A true positive is an observed functional residue (as defined in the coordinate entry) inside the predicted site, a false positive is a residue inside the predicted site that is not defined as forming the observed functional site (over prediction) and, finally, a false negative is an observed functional residue not included in the predicted site (under prediction).

The residue sensitivity reports how many observed functional residues have been predicted correctly and has an average value of 0.72. The residue specificity reports how many of the residues predicted to be functional really do form the functional site, and the average value is 0.32. The low level of specificity is not surprising, as we have generated a sphere around the predicted functional site centroid, and generally inside this sphere there are both the catalytic residues and the surrounding ones. To assess this effect, the specificity was recalculated using only the functional site residues as defined in the PDB to generate a sphere (based on calculating the centroid and using a radius as the largest distance from this centroid to the observed C^{β} atom). In this ideal prediction, the average specificity was 0.45, which identifies the average of the upper limit that could be obtained.

Assessing Inheritance of Protein Function

Strategy

The problem considered is that during genome annotation, two proteins (A and B) are identified as homologues. The three-dimensional structure of one protein (or of its close homologue) is known (protein A). In addition, the function for one (either A or B) is known and one wishes to assess whether the other protein has a similar or a quite different function. For simplicity, we now assume that protein A has known function. Here, we propose the following strategy.

Step (i), apply the functional site prediction starting with protein A. Terminate the construction of the phylogenetic tree just before the sequence of B (along with any of its close homologues) are added. Generally, one or more functional sites will be predicted for the family A and one proceeds as follows. Step (ii), now include the sequence of B (with the sequences of its close homologues) into the phylogenetic tree and rerun the functional site algorithm. If one still gets a functional site prediction, then the approach predicts that the proteins have a related function. However, if no functional site can be predicted, the approach suggests that A and B have quite different functions.

Occasionally, no site is predicted by step (i), either because there are not even two invariant polar residues or, if there are two (or more) polar residues, they do not cluster spatially. In this situation one cannot proceed with this approach.

To evaluate this proposed strategy, we used the table of 19 pairs of homologous proteins with different function reported by Hegyi & Gerstein.[4] This set includes homologous pairs that are enzyme and non-enzyme, and pairs that are both enzymes but with quite different functions. In addition, to evaluate the counter case when proteins A and B have similar function, we examined the list of 86 proteins for which the phylogenetic tree was able to attain 30 % or lower identity. For each of these proteins, all homologues in the multiple alignment must have similar functions to provide the counter case. By inspection of the names and EC numbers of the clustered sequences, we identified 16 proteins that spanned two or more functional families and these were removed from the set of proteins with common function to yield 70 proteins.

Before presenting the results on the above data sets, we first consider what would happen if the functional site prediction were applied to these 16 proteins that we needed to exclude to establish the dataset of counter cases. A key feature of the functional site prediction algorithm is that it restricts the diversity of the proteins included in the alignment by ensuring that there is at least one predicted functional site. For 11 of these 16 proteins, the prediction algorithm (above) would have used the appropriate level of attained identity and there-

Table 2. Accuracy of the assessment of functional inheritance

	No. of proteins	No. of predicted as related function	No. of predicted as different function	
			Sequence information	Structure information
Proteins with related function	70	68	-	2
Proteins with different function	18	4	9	5

The columns give the total number of proteins with related and unrelated function used in the study, the number of proteins predicted as having a related function, and the number of proteins predicted as having an unrelated function using only sequential information, and structural information respectively.

by would have correctly identified a common function for all proteins included in the alignment. In contrast, for five of these 16 proteins, too low a level of attained identity would have been used and a common function would have been predicted erroneously.

Results

Table 2 summarises the results of this strategy. The algorithm could be applied to 18 of the 19 pairs of protein listed by Hegyi & Gerstein.[4] The algorithm could not be applied to one of the pairs ENTA_ECOLI and ADHI_DROMO. This is because no functional site could be predicted for ENTA_ECOLI due to the multiple alignment yielding no invariant residues. Inspection of the descriptions of the aligned sequences for ENTA_ECOLI did not suggest that there were multiple functions for this family and thus the failure to identify invariant residues is a limitation of the algorithm.

The method correctly predicted 68 out of the 70 pairs as having related functions. For the 18 pairs with different functions, 14 were correctly predicted as having different functions while four pairs were incorrectly assigned as having related functions (Table 3). For nine of these 14 pairs, the assignment was based simply on the lack of, at least, two invariant polar residues and thus the prediction could have been made from sequence considerations alone. But for the remaining five pairs, one required the three-dimensional information as used in the spatial clustering. This illustrates the important role of three-dimensional information in assigning function to sequence.

Overall, the accuracy of correct predictions divided by the total number of pairs considered is 94 % (82/88). Since there are different numbers of proteins in the dataset with similar and different functions, the accuracy of a random prediction is not 50 %. The best possible random prediction would simply be to predict that all pairs have the same function (the largest class prediction) and the corresponding random accuracy would then be 79 % (70/88). The difference between 94 % and 79 % accuracy is significant at better than 0.001 using a χ^2 test on the 2×2 contingency table. Note that this random prediction, although achieving 79 % accuracy, would not identify any pair of proteins as having different functions.

The success of the prediction is not a consequence of having greater sequence identities between the homologues with related function than between the homologues with different function. The attained percentage identities once all the homologous proteins have been aligned are similar, and independent of the function (data not shown).

Figure 3 illustrates the methodology on 2abk_ (endonuclease III, EC 4.2.99.18) in which the enzyme active site is identified correctly. When the possible G-T mismatch repair enzyme (EC 3.2.2.-) is included in the functional site generation, no site is predicted. One infers correctly that the two enzymes have different activity. When the two proteins are joined in their multiple sequence alignment, there still remain seven invariant polar residues but they do not form a spatial cluster.

We now consider why the algorithm failed for five protein pairs (Table 3). For SCOP code 1bdo_, the proteins are biotin carboxyl carrier protein of acetyl-CoA carboxylase (EC 6.4.1.2) and biotin carboxyl carrier protein of methylmalolyl-CoA carboxyl transferase (EC 2.1.3.1). Although the proteins have different EC numbers, they are both biotin carriers and the functional site prediction correctly identifies the common biotin-binding residues. For 1nipA, nitrogenase iron protein (EC 1.18.6.1) and protochlorophillide reductase 33 kDa subunit (EC 1.3.1.33), the enzymes share a similar ATP-binding site that is predicted correctly by the algorithm. For 1dhpA, dihydrodipicolinate synthase alcohol dehydrogenase 1 (EC 4.2.1.52) and *N*-acetylneuraminate lyase subunit (EC 4.1.3.3), both proteins use the same catalytic residue Lys164, which was identified correctly in the predicted functional site. For 1isu, no homologue to either of the two proteins was found in the sequence database, and this led to a 29 % attained identity that hindered the algorithm.

These observations highlight the problems of the inheritance of function. There were two pairs of proteins that had a common binding site for one moeity but different catalytic functions. In both cases, the algorithm predicted that the pairs had similar function. Thus, in practice, if the protein of known function has a set of subfunctions, such as

Table 3. Application of functional sites method to identify homologues with different function

	Function 1					Function 2					Function 1+2	
SCOP domain	Swisprot code	EC number	Function	%ID	Number of funct. sites	Swisprot code	EC number	Function	%ID	Number of funct. sites	%ID	Number of funct. sites
Two different enzymatic functions												
2abk_	END3_ECOLI	4.2.99.18	Endonuclease III	4.37	1	GTMR_METTF	3.2.2.-	Possible G-T mismatches repair enzyme	100	14	1.56	-(*)
1bdo_	BCCP_ECOLI	6.4.1.2	Biotin carboxyl carrier of acetyl-CoA carboxylase	20.0	1	BCCP_PROFR	2.1.3.1	Biotin carboxyl carrier of methylmalolyl-CoA carboxyl-transferase	8.33	1	1.61	1
1dhpA	DAPA_TACSU	4.2.1.52	Dihydrodipicolinate synthase, alcohol dehydrogenase 1,	8.79	3	NPL_ECOLI	4.1.3.3	*N*-acetylneuraminate lyase subunit	55.88	15	2.56	1
1hdcA	ENTA_ECOLI	1.3.1.28	2,3-dihydro-2,3 dihydroxy-benzoate dehydrogenase	99.6	20	ADHI_DROMO	1.1.1.1	Alcohol dehydrogenase 1	100	20	2.08	N/A
1nipA	NIFH_THIFE	1.18.6.1	Nitrogenase iron protein	24.83	5	bCHL_RHOCA	1.3.1.33	Protochlorophillie reductase 33 kDa subunit	42.42	8	8.53	1
1garA	PUR3_YEAST	2.1.2.2	Phosphoribosyl glycinamide formyltransferase	2.41	-	PURU_CORSP	3.5.1.10	Formyltetrahydrofolate deformylase	15.46	3	1,71	-
2dkb_	OAT_RAT	2.6.1.13	Ornithine aminotransferase precursor	73.38	26	GSAB_TACSU	5.4.3.8	Glutamate-1-semialdehyde 2,1-aminomutase 2	97.05	25	2.43	- (*)
1ede_	HALO_XANAU	3.8.1.5	Haloalkane dehalogenase	8.72	2	DMPD_PSEPU	3.1.1.-	2-Hydroxymuconic semialdehyde hydrolase	0.25	-	0.24	-
1fua_	FUCA_ECOLI	4.1.2.17	L-Fuculose phosphate aldolase	39.13	5	ARAD_ECOLI	5.1.3.4	L-Ribulose 5-phosphate 4-epimerase	15.35	1	0.43	-
1lmn_	LYC1_PIG	3.2.1.17	Lysozyme C-1	22.31	2	LCA_RAT	2.4.1.22	Alpha-lactalbumin precursor	63.16	7	7.46	-
1frvA	MBHS_AZOCH	1.18.99.1	Uptake hydrogenase small subunit precursor	2.3	3	FRHG_METVO	1.12.99.1	Coenzyme F420 hydrogenase gamma subunit	0.94	-	0	-
Enzyme and non-enzyme												
1gsq_	GTS2_MANSE	2.5.1.18	Glutathione *S*-transferase 2	100	22	SC11_OMMSL		*s*-cRYSTALLIN sl11	27.06	-	19.27	- (*)
1lcl_	lPPL_HUMAN	3.1.1.5	Eosinophil lysophospholipase	36.17	5	LEG7_RAT		Galectin-7	26.2	-	1.41	-
1brbE	CFAD_RAT	3.4.21.46	Endogenous vascular elastase	4.81	1	CAP7_HUMAN		Azurocidin	12.81	-	1.86	-
1mup_	PGHD_HUMAN	5.3.99.2	Prostaglandin-D synthase	20.75	4	LACC_CANFA		Beta-lactoglobulin III	26.75	4	0	-
..1mup_						QSP_CHICK		Quiescence-specific protein	100	16	0	-
2hhmA	MYOP_XENLA	3.1.3.25	Inositol mono-phosphatase	16.73	2	SUHB_ECOLI		Extragenic suppressor protein	28.06	2	5.98	- (*)
..2hhmA	STRO_STRGR	2.7.7.24	DTDP-glucose synthase	100	14						1.52	- (*)
1isuA	IRO_THIFE	1.16.3.-	Iron oxidase precursor	100	7	HPIT_RHOTE		High potential iron-sulfur protein	100	7	29.03	1

A list of SCOP domains that are homologous to several proteins with significantly different functions as given by Hegyi & Gerstein.[4] The first column gives the SCOP domain code. The following columns give the SwissProt code, EC number, function description, %ID attained in the multiple alignment and the number of functional sites predicted for each protein function described independently. The last two columns give the %ID attained when the different families are clustered together in the multiple alignment and the final number of functional sites predicted. The asterisk shows the cases when the different functions cannot be identified without structural information.

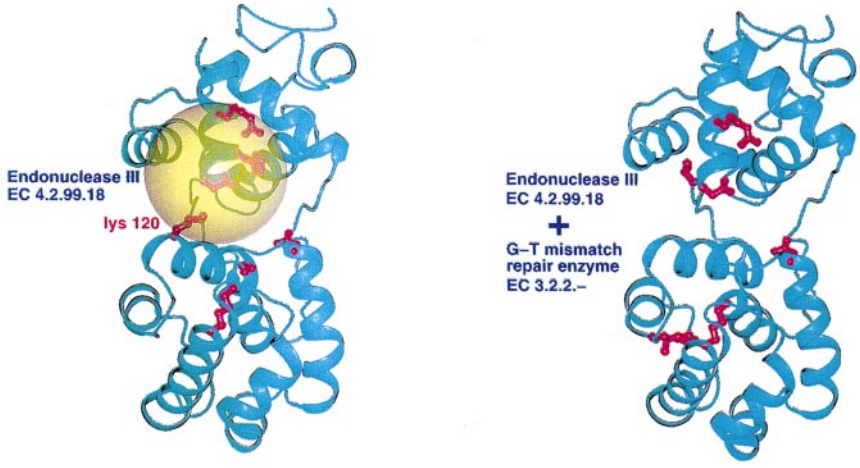


Figure 3. Example of functional divergent evolution for endonuclease III and G-T mismatch repair enzyme families (2abk SCOP domain). All the conserved residues for the endonuclease III family are represented as ball-and-stick. The predicted functional site is shown in yellow. Once the two different families have been clustered together in the multiple alignment, and despite some remaining conserved polar residues, the predicted functional site disappears.

both a catlytic activity and a binding site, and the algorithm predicts common function, one should infer that maybe only one of the subfunctions will be inherited. Further studies examing other family members could assist in deciding which of the subfunctions should be transferred. In general, we note that the approach is marginally better at distinguishing enzyme from non-enzyme (seven out of eight correctly identified) than between enzymes of different activity (seven out of ten correctly identified).

Application to Macromolecular Docking

Several groups are developing algorithms to predict the structure of protein-protein and protein-DNA complexes starting from their unbound components (e.g. see references[24,25,36–41]), see Sternberg *et al*.[42] for a review. The docking algorithms identify several solutions and generally only one (or a few) of these are close to the true complex. These solutions are then ranked by a score generated by the docking program or by a subsequent scoring scheme. One evaluates the highest rank at which a close solution to the native complex is found. To reduce the size of the list of possible solutions, biological information about residues involved in the macromolecular recognition is often used. Typical biological information that can be used includes known active-site residues or the results of mutagenesis. However, sometimes the biological information is not available and then a method for predicting the regions involved in recognition is required. Here, we evaluate the filtering power that could be obtained automatically using functional site prediction and compare the results to those obtained manually using biological information. Note that here we consider only the first two steps of the docking strategy: the generation by FTDOCK[25] of a set of complexes with favourable shape and electrostatic complementarity, and the use of distance constraints. A further filtering of the list of possible complexes can be obtained by using empirical residue-residue or residue-DNA pair potentials[24,27] and for refinement of the interface of protein-protein complexes.[26]

We examined two sets of systems that we have studied previously: five complexes for serine proteinases docking to inhibitors[25] and seven complexes for repressors binding to DNA.[24] The list of complexes was generated using the program FTDOCK.[25] Functional site predictions were run for each serine protease and for each repressor. The distance constraint filter requires that at least one of the predicted functional residues is close to any part of the other molecule, otherwise the complex fails to pass the filter and is removed from the list. The precise distant constraints were those used in the earlier docking studies (and see the legend to Table 4).[24,25]

Table 4 shows that the results using predicted functional sites are very similar to those using manually introduced experimental information, in particular, active-site information for enzymes and the results of mutagenesis for repressors. The benchmarking of the functional site prediction (see above) shows that the approach works best when the multiple sequence alignments achieve $\leqslant 30\,\%$ attained identity. All the enzyme-inhibitor complexes are serine proteases and, consequently the algorithm can find a large number of homologous proteins in the databases, which allows us to attain low identity levels in the multiple alignments (10-20 %). For three out of five complexes, the results are identical between using predicted functional sites and the experimental active site. This is the consequence of the predicted functional site being identical with the experimental information: we

Table 4. Use of functional sites in protein docking

Complex	Range of %ID	% Overlap	Rank of the first good solution		
			Prior filtering	Funct. sites information	Experimental information
A. *Protein-protein*					
CGI	⩽10	80	87	2	2
CHO	⩽10	75	127	2	2
KAI	20-10	67	223	25	45
PTC	20-10	79	502	9	3
SNI	⩽10	100	1518	2	2
B. *Protein-DNA*					
ARC	50-40	58	91	41	22
CRO	40-30	87	12	3	3
GAL	100-90	-	37	-	26
LAC	100-90	82	30	13	13
LAM	30-20	67	22	3	3
PUR	50-40	100	9	2	2
TRP	90-80	40	4	2	2

Docking results for the protein-protein and protein-DNA complexes starting from unbound coordinates. The first column gives the complex code identifier, the second the %ID attained in the multiple alignment and the third the percentage of overlap between the real complex interface and the predicted functional sites. The fourth column gives the rank of the first near-native complex predicted by FTDOCK prior to filtering and after filtering, using functional sites and experimental information respectively. The distance constraints are those used in the original docking papers.[24,25] For protein-protein complexes, the distance constraint (D) is defined between the C^{α} atom of any residue inside the predicted functional site and the C^{α} atom of any residue of the inhibitor and D must be less than (L+ 4 Å), where L is the specific side-chain length (ranging from 0.5 Å for Gly to 6.0 Å for Arg). For protein-DNA complexes, D is defined between the C^{α} atom of any residue inside the predicted functional site to any N1′ of the DNA footprint. A complex passes the filter when $D(C^{\alpha}\text{-}N1') < L + 4$ Å.

have predicted the catalytic triad. However, for trypsin with bovine pancreatic inhibitor (PTC complex), the predicted filtering is poorer than using experimental data, the rank of the first good solution drops from position 3 (experimental) to 9 (predicted), although we are still raising it from position 502 before any filter was applied. This poorer filtering is because two different functional sites were identified instead of predicting only the active site. In contrast, the results for kallikrein bound to bovine pancreatic trypsin inhibitor (KAI complex) are better using the fully automatic method, raising the rank of the first good complex from 45 to 25. This was because in the crystal structure the residues of the catalytic triad are a slightly further apart than in the other proteinases and the functional site method clustered only two of the three active residues, and this proved to be a superior filter. Figure 4(a) shows how the predicted functional site matches perfectly to the known α-chymotrypsin active site and how it is used for filtering the docking complexes.

For the repressor-DNA complexes, the multiple sequence alignments do not achieve the desired level of ⩽30 %, ranging from just over 30 % up to 100 %. This is due, in part, to the fewer homologous sequences in the databank. Nevertheless, the information obtained from the predicted functional sites was still useful and the results were generally similar to those obtained when experimental information was used. The repressors used in this work are mainly small proteins that often form dimers

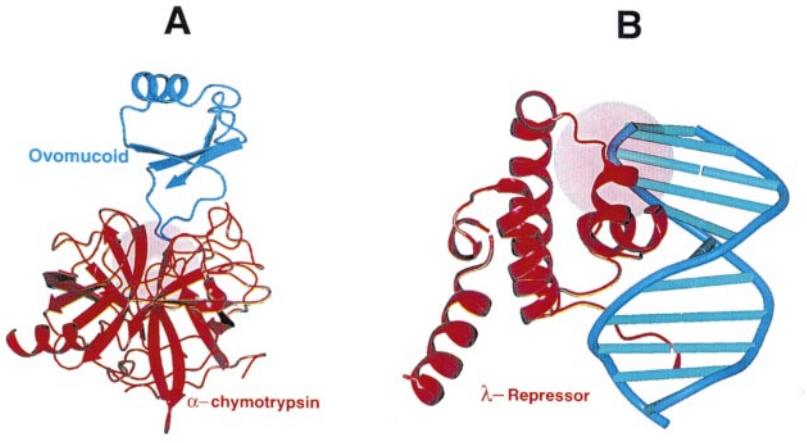


Figure 4. Examples of the automatic application of the functional site prediction as a geometric filter for protein-protein and protein-DNA docking. The First good solutions generated by FTDOCK for (a) α-chymotrypsin and (b) lambda repressor are shown.

or larger homocomplexes when binding DNA. Consequently, most of the protein surface is used for binding either DNA or the other subunits. The DNA-binding patch tends to use polar residues for the nucleic acid recognition, whilst the subunit-binding region involves more hydrophobic interactions. Since we are clustering in the space only polar residues to define the functional sites, this helps to identify the polar DNA-binding site. For five out of the seven repressor-DNA complexes analysed, the rank of the first good solution is identical with that with the interacting residues defined manually.

ARC is a small dimer that uses a two-stranded antiparallel β-sheet to recognise the major groove of the DNA. The method was applied to only one of the two chains of the dimer, and when the functional sites information was used as a filter, a large functional site from the combined two chains was generated and this yielded poor filtering power. The rank obtained was half of that reached prior to filtering (rank 91 → rank 41), which is poorer than when experimental information was used (rank 22).

There are two systems (LAC and GAL) for which no homologous protein could be found in the databanks (i.e. 100 % attained identity). Despite this, for the LAC repressor, clustering the polar residues successfully identified the DNA-binding site and the filtering was as powerful as using experimental information. In contrast, for GAL, the predicted functional sites were completely wrong and once this information was automatically used as a filter, the method removed the good solutions from the list generated by FTDOCK.

Figure 4(b) shows the first native-like predicted complex (rank 3) for lambda repressor. Despite the helix-turn-helix DNA binding motif not being identified perfectly, some of the recognition residues are identified, and this information is enough to permit the removal of 748 false solutions and to raise the first good prediction from rank 22 to rank 3.

Discussion and Concluding Remarks

This study has shown that one can automate the prediction of functional sites in proteins. The evolutionary family must be sufficiently diverse so that the percentage of invariant residues is $\leqslant 30$ %. For these protein families, 51 % of the predicted functional sites will overlap by at least 50 % in volume with the observed active site. For 79 % of the proteins, the method will yield some information about the functional residues. In contrast to the work of Lichtarge *et al.*,[18–20] we have not considered predicting the residues important for specificity within a subfamily of the aligned sequences. This would form the topic for further investigation extending the sequence approach used by Hannehalli & Russell[17] by including structural data.

The approach is based on clustering invariant polar residues and clearly can be applied only to proteins where such residues form the functional site. Our data set included 95 enzymes and 11 non-enzymes. According to the PDB definition of functional residues, only two proteins had no polar residues in the functional site and seven had only one. In addition to these PDB-defined functional residues, our inspection showed that, generally, there were other invariant polar residues in the vicinity of the functional site. These observations would be expected for enzymes, since polar residues form the active site. However, for other types of proteins, such as those involved in protein-protein interactions, there still remain invariant polar residues mediating the interaction. Jones & Thornton[43] analysed protein-protein interactions from a wide range of systems and consistently observed the presence of polar residues in the interfaces, since the component proteins must be independently stable. Indeed, the results of the functional site prediction worked well for both the enzymes and the non-enzymes in the data set. Furthermore, the algorithm was used successfully in the docking application on both enzyme-inhibitor and repressor-DNA complexes.

The success of the prediction of functional sites depends on having identified invariant residues correctly, and this requires an accurate multiple sequence alignment. The lowest pairwise identities used to construct the alignments ranged from 21 % to 75 % with a mean of 34 %. The accuracy of the multiple alignments generated by several programs, including CLUSTALV, the predecessor of CLUSTALW, has been evaluated by McClure *et al.*[44] They evaluated the accuracy of identifying local conserved motifs that will often include the invariant residues used in our algorithm. The conclusion from their study was that even for alignments with sequences with 20-30 % identity, conserved local motifs can be aligned reliably. Thus, in general, errors in sequence alignment are not a problem for this automated approach.

The method would also be sensitive to sequence errors. In the approach, a single residue in one sequence is allowed to be different from the remaining invariant residue type at a location. However, even if there were more than one sequence error at a position, then the algorithm might still work. There could be more than two polar residues that could be clustered to form the predicted site and then clustering could still work even without all the residues included. Nevertheless, sequence errors or point mutations at some of the active-site residues can lead to difficulties.

The automated method of predicting functional sites can provide a useful guide to assess whether two homologous proteins have a related function. The approach is based on the principle that if the two proteins have different functions, one would be unable to predict a common functional site. The

study considered proteins with around 30 % pairwise identity and unrelated functions. The approach is powerful, with 82 out of the 88 test pairs being assigned correctly. About one-third of the correct predictions of different function between pairs of proteins exploited three-dimensional structural information. The approach can distinguish between homologous proteins with different functions, both between enzymes and non-enzymes and between different families of enzymes.

To illustrate the implications for genome annotation, we have considered the 4257 proteins (1,351,301 residues) in the *Escherichia coli* genome.[45] Work in progress at the ICRF (A. Mueller, R. MacCallum & M.J.E.S.) shows that using PSI-BLAST,[46] 11 % of the residues in *E. coli* are in regions that have a homology with $\leqslant$25 % identity to another sequence with a clear functional annotation from the sequence databases. This 25 % cut-off is used to identify the level below which the transfer of function *via* homology is unreliable.[5,6] Of these 11 % of the residues, 40 % are in regions that are homologous to a known structure and therefore amenable to the functional prediction. Thus, at present, our approach to prevent incorrect functional transfer could be applied to 4 % of a bacterial genome and this coverage will increase as more structures are determined.

The prediction of functional sites to act as filters in a predictive scheme for protein-protein and protein-DNA docking was shown, in general, to be as effective as manually introducing biological constraints of the type that typically would be available in many docking applications. Thus, functional site prediction has a central role in the development of a fully automated strategy for macromolecular docking.

With genome sequencing and structural genomics initiatives, the automated approach described here should have many applications to provide functional information from protein structure and sequence. Work is in progress to offer this methodology to the scientific community as a web server.

Materials

Protein data set

The Protein Data Bank codes of the non-redundant database used are given below. The last character corresponds to the chain identifier, with _ denoting none.

1abrA, 1acbI, 1af0A, 1ah7_, 1ai7B, 1aihB, 1aj8A, 1al6_, 1aop_, 1ast_, 1at0_, 1auoA, 1ayl_, 1bam_, 1bcrB, 1bdb_, 1bhs_, 1bia_, 1blsA, 1bplB, 1bpxA, 1broA, 1burA, 1cglA, 1chd_, 1chkA, 1cleA, 1clxA, 1cmvB, 1cug_, 1cvl_, 1dtp_, 1dxy_, 1esb_, 1eur_, 1exfA, 1fpgA, 1fwcC, 1gcb_, 1glm_, 1har_, 1hny_, 1hpgA, 1huh_, 1idk_, 1inf_, 1jud_, 1kit_, 1ksiA, 1lbu_, 1lml_, 1lpbB, 1masA, 1mat_, 1meg_, 1mkbA, 1mla_, 1mra_, 1mvpA, 1niaA, 1nje_, 1odwA, 1onc_, 1pnt_, 1pvaA, 1qca_, 1quf_, 1rne_, 1rtu_, 1sacA, 1se3_, 1sgt_, 1skyE, 1sphA, 1sta_, 1tca_, 1tpb1, 1uch_, 1udg_, 1vdc_, 1vsb_, 1vsj_, 1wab_, 1xgsA, 1xjo_, 1xpb_, 1xtcA, 1yasA, 1zap_, 2ace_, 2acr_, 2btfA, 2dubD, 2ebn_, 2hntE, 2myr_, 2pcdA, 2pcdM, 2sim_, 2tmy_, 3bir_, 3lz2_, 3ptd_, 3pte_, 3tgl_, 4enl_, 4fua_, 5csmA, 6acn_, 6rsa_, 7cel_.

The representative dataset used a 25 % identity cut-off. At this cut-off, there will be cases where a homologue is excluded but could have been used in the benchmark as it had a function different from that of the representative protein that was included. However, if we had used a higher identity (say 40 %), then we would have generated a highly biased database that included many homologues with similar function. In a benchmark, it is clearly better to exclude some proteins than to introduce a marked bias due to many homologues and therefore we used this 25 % identity cut-off.

Dependency on parameters

C^{β} atoms were used for clustering as they provide an appropriate representation of the location of the side-chains that often are involved in the function of the protein. Indeed, if C^{α} atoms rather than C^{β} atoms are used for clustering, then the number of observed functional sites identified correctly falls from 68 to 52. With C^{β} clustering, the precise choice of the two distance cut-offs for clustering is not critical to the resultant accuracy of predicting the observed functional sites (ranging from 59 to 69 sites identified correctly for $\pm$1 Å in the two cut-off distances).

Acknowledgements

The authors thank Jaap Heringa (NIMR, London) for help with the OBSTRUCT program, Arne Mueller (ICRF) for data on the *E. coli* genome and Peter Sasieni (ICRF) for statistical advice. P. A. is a fellowship recipient from the CICYT (Ministerio de Educación y Cultura, Spain). This work has been supported by grants BIO2000-0647 and BIO98-0362 from the CICYT, by CERBA (Centre de Referència de Biotecnología de la Generalitat de Catalunya), by C4-CESCA (Barcelona, Spain) and by the Imperial Cancer Research Fund.

References

1. Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. & Yuan, Y. (1998). Predicting function: from genes to genomes and back. *J. Mol. Biol.* **283**, 707-725.
2. Bork, P. & Koonin, E. V. (1998). Predicting functions from protein sequences - where are the bottlenecks. *Nature Genet.* **18**, 313-332.
3. Russell, R. B., Sasieni, P. D. & Sternberg, M. J. E. (1998). Supersites within superfolds - binding site similarity in the absence of homology. *J. Mol. Biol.* **282**, 903-918.
4. Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147-164.
5. Wilson, C. A., Kreychman, J. & Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297**, 233-249.

6. Devos, D. & Valencia, A. (2000). Practical limits of function prediction. *Proteins: Struct. Funct. Genet.* **41**, 98-107.
7. Doerks, T., Bairoch, A. & Bork, P. (1998). Protein annotation: detective work for function prediction. *Trends Genet.* **14**, 248-250.
8. Brenner, S. E. (1999). Errors in genome annotation. *Trends Genet.* **15**, 132-133.
9. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. L. (2000). The Pfam protein families database. *Nucl. Acids Res.* **28**, 263-266.
10. Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. & Bork, P. (2000). SMART: a web-based tool for the study of genetically mobile domains. *Nucl. Acids Res.* **28**, 231-234.
11. Bairoch, A. & Apweiler, R. (2000). The SWISS-Prot protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45-48.
12. Henikoff, J. G., Greene, E. A., Pietrokovski, S. & Henikoff, S. (2000). Increased coverage of protein families with the Blocks Database servers. *Nucl. Acids Res.* **28**, 228-230.
13. Lo Conte, L., Ailey, B., Hubbard, T. J. P., Brenner, S. E., Murzin, A. G. & Chothia, C. (2000). SCOP: a structural classification of proteins database. *Nucl. Acids Res.* **28**, 257-259.
14. Webb, E. C. (1992). *Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*, Academic Press, New York.
15. Casari, G., Sander, C. & Valencia, A. (1995). A method to predict functional residues in proteins. *Nature Struct. Biol.* **2**, 171-178.
16. Livingstone, C. D. & Barton, G. J. (1996). Identification of functional residues and secondary structure from protein multiple sequence alignment. *Methods Enzymol.* **266**, 497-512.
17. Hannenhalli, S. S. & Russell, R. B. (2000). Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.* **303**, 61-76.
18. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). Evolutionarily conserved G alpha beta gamma binding surfaces support a model of the G protein-receptor complex. *Proc. Natl Acad. Sci. USA,* **93**, 7507-7511.
19. 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.
20. 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.
21. Lockless, S. W. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science,* **286**, 295-299.
22. Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D. & Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners. *J. Mol. Biol.* **299**, 283-293.
23. 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.
24. Aloy, P., Moont, G., Gabb, H. A., Querol, E., Aviles, F. X. & Sternberg, M. J. E. (1998). Modeling repressor proteins docking to DNA. *Proteins: Struct. Funct. Genet.* **33**, 535-549.
25. Gabb, H. A., Jackson, R. M. & Sternberg, M. J. E. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.* **272**, 106-120.
26. Jackson, R. M., Gabb, H. A. & Sternberg, M. J. E. (1998). Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *J. Mol. Biol.* **276**, 265-285.
27. Moont, G., Gabb, H. A. & Sternberg, M. J. E. (1999). Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins: Struct. Funct. Genet.* **35**, 364-373.
28. 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.
29. Heringa, J., Sommerfeldt, H., Higgins, D. & Argos, P. (1992). OBSTRUCT: a program to obtain largest cliques from a protein sequence set according to structural resolution and sequence similarity. *Comput. Appl. Biosci.* **8**, 599-600.
30. Hobohm, U., Sander, C., Scharf, M. & Schneider, R. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409-417.
31. Holm, L. & Sander, C. (1998). Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics,* **14**, 423-429.
32. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.
33. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1998). Assessing sequence comprison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA,* **95**, 6073-6078.
34. Thompson, J. D., Higgins, D. G. & Gibson, J. J. (1994). CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673-4680.
35. Everitt, B. (1974). *Cluster Analysis*, Heineman, London, chapt. 3.
36. Vakser, I. A. (1997). Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. *Proteins: Struct. Funct. Genet.* **1**, 226-230.
37. Norel, R., Petrey, D., Wolfson, H. J. & Nussinov, R. (1999). Examination of shape complementarity in docking of *unbound* proteins. *Proteins: Struct. Funct. Genet.* **36**, 307-317.
38. Ritchie, D. W. & Kemp, G. J. L. (2000). Protein docking using spherical polar Fourier correlations. *Proteins: Struct. Funct. Genet.* **39**, 178-194.
39. Knegtel, R. M. A., Antoon, J., Rullmann, C., Boelens, R. & Kaptein, R. (1994). MONTY: a Monte Carlo approach to protein-DNA recognition. *J. Mol. Biol.* **235**, 318-324.
40. Shoichet, B. K. & Kuntz, I. D. (1991). Protein docking and complementarity. *J. Mol. Biol.* **221**, 327-346.
41. Ausiello, G., Cesareni, G. & Helmer-Citterich, M. (1997). ESCHER: a new docking procedure applied to the reconstruction of protein tertiary structure. *Proteins: Struct. Funct. Genet.* **28**, 556-567.
42. Sternberg, M. J. E., Gabb, H. A. & Jackson, R. M. (1998). Predictive docking of protein-protein and protein-DNA complexes. *Curr. Opin. Struct. Biol.* **8**, 250-256.

43. Jones, S. & Thornton, J. M. (1997). Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121-132.
44. McClure, M. A., Vasi, T. K. & Fitch, W. M. (1994). Comparitive analysis of multiple protein-sequence alignment methods. *Mol. Biol. Evol.* **11**, 571-592.
45. Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V. & Riley, M. *et al*. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science,* **277**, 1453-74.
46. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein data base search programs. *Nucl. Acids Res.* **25**, 3389-3402.

Edited by G. von Heijne

(Received 22 January 2001; received in revised form 16 May 2001; accepted 21 June 2001)