Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W89–W93***
doi:10.1093/nar/gki414

ProFunc: a server for predicting protein function from 3D structure

Roman A. Laskowski*, James D. Watson and Janet M. Thornton

European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received February 15, 2005; Revised and Accepted March 21, 2005

ABSTRACT

ProFunc (http://www.ebi.ac.uk/thornton-srv/databases/ProFunc) is a web server for predicting the likely function of proteins whose 3D structure is known but whose function is not. Users submit the coordinates of their structure to the server in PDB format. ProFunc makes use of both existing and novel methods to analyse the protein's sequence and structure identifying functional motifs or close relationships to functionally characterized proteins. A summary of the analyses provides an at-a-glance view of what each of the different methods has found. More detailed results are available on separate pages. Often where one method has failed to find anything useful another may be more forthcoming. The server is likely to be of most use in structural genomics where a large proportion of the proteins whose structures are solved are of hypothetical proteins of unknown function. However, it may also find use in a comparative analysis of members of large protein families. It provides a convenient compendium of sequence and structural information that often hold vital functional clues to be followed up experimentally.

INTRODUCTION

A large proportion of the structures deposited at the PDB (1) by the various structural genomics initiatives (2) are of 'hypothetical proteins', i.e. proteins of unknown function. These are classed as hypothetical when sequence search methods have failed to match them to proteins that have been functionally characterized. However, knowing the 3D structure of a protein opens up the possibility of ascertaining its function from an analysis of that structure. Recently, many methods have been developed for predicting protein function from structure. These range from global comparisons, such as matching the protein's fold against other proteins of known 3D structure, to identification of more local features, such as active site residues or DNA-/ligand-binding motifs (3). None of these structure-based methods can expect to be successful in all cases. For example, methods that are able to detect catalytic residues in a 3D structure will give no useful information if the protein in question is not an enzyme. Therefore, a prudent approach is to use as many methods as possible, both structure-based and sequence-based, not only to increase the chances of obtaining a helpful match, but also to benefit from cases where several methods arrive at the same or similar conclusions.

This is the principle behind the ProFunc server (4), which runs a number of different methods to analyse both the sequence and the structure of a submitted protein and provide a single, convenient summary of what each method has found.

THE ProFunc SERVER

Figure 1 shows the analyses that are currently run whenever a protein structure is submitted to ProFunc (http://www.ebi.ac.uk/thornton-srv/databases/ProFunc).

Sequence-based searches

The first batch of processes, on the left-hand side of Figure 1, are all standard sequence-related searches. Two runs of Blast (5) are made: the first against the protein sequences with known structures in the PDB to find any obvious matches to proteins of known structure, and the second against the UniProt database of protein sequences (6). The results from the latter are aligned using a simple pile-up procedure to give a multiple sequence alignment, from which residue conservation scores are computed for the target sequence using the method described previously by Valdar and Thornton (7). Residue conservation plays a key part in some of the structural analyses to be described below.

For every UniProt sequence matched by Blast, the protein's location on its genome is found and its 10 gene neighbours on either side are extracted. These are tabulated and illustrated in a schematic diagram. Neighbouring genes are often functionally related, so a functionally characterized neighbour may

*To whom correspondence should be addressed. Tel: +44 1223 492 542; Fax: +44 1223 494 468; Email: roman@ebi.ac.uk

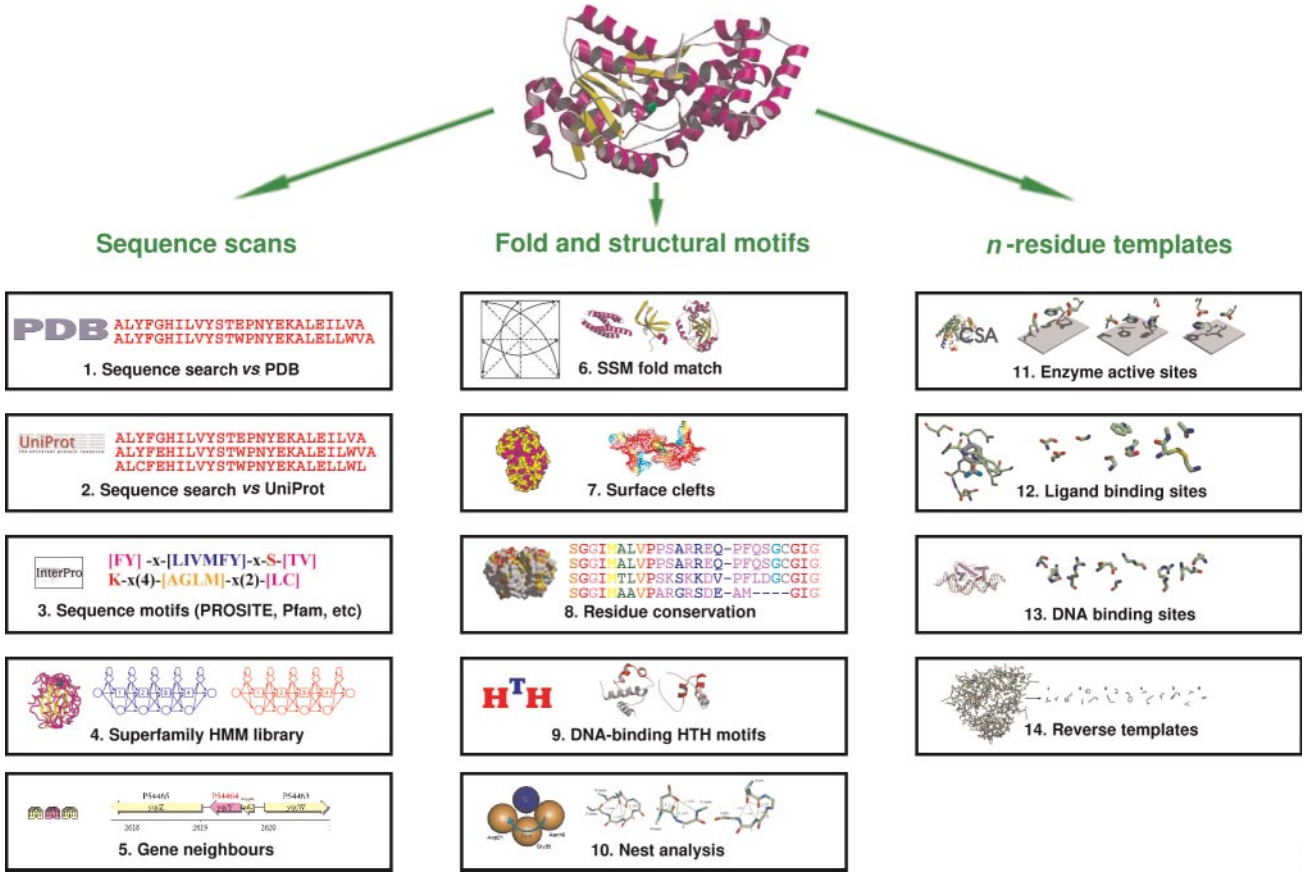

Figure 1. Schematic diagram showing the different methods, both sequence- and structure-based, that are applied to a given structure when it is submitted to the ProFunc server. The leftmost column shows the methods that use the sequence of the submitted protein, the middle column lists the methods that analyse the protein's structure, and the rightmost column lists the 3D template methods that are used to match the structure to existing PDB entries.

provide some clue to the target protein's role. Where the target protein's own genome has not been determined, or is not informative, the functional information can sometimes come from the genome of one of the other sequences matched by Blast. Figure 2 shows an example.

Finally, the target sequence is scanned against the numerous motifs, patterns, fingerprints and hidden Markov models (HMMs) of Interpro (8) by running InterProScan (9). InterPro holds the sequence patterns from the PROSITE, Pfam, SMART, PRINTS, BLOCKS, TIGR and ProDom databases. A separate scan is then performed against the SUPERFAMILY (10) library of HMMs derived from the SCOP structural superfamilies, potentially giving matches to the existing PDB entries or parts thereof.

Structure-based analyses

The second batch of processes that the target structure undergoes is the set of structure-based analyses listed in the middle column of Figure 1. The first of these is a search for known structures having the same, or similar, overall fold as the target. The program used is secondary structure matching (SSM) (11), which uses a fast graph-matching algorithm to compare the secondary structure elements (SSEs) of the target structure against those of the structures in its database. Any strong matches are superposed and an r.m.s.d. for equivalent C-alphas is calculated. Finding a fold relative can often provide strong indications about the target protein's likely functional type.

The SURFNET algorithm (12) is then used to compute all the clefts in the protein structure, ranking them in order of size. The clefts can be viewed in RasMol (13) via automatically generated scripts that colour the cleft surfaces by specific properties, such as cleft size, residue type or conservation score. Residue conservation is a particularly powerful means of highlighting which are the key residues in the structure, and so can usually help pick out the most likely location of the protein's functional site(s) (14–16). The residue conservation scores are also mapped onto the whole protein surface, which, again, can be viewed in RasMol. This can help identify other functionally important regions in the structure, such as likely protein–protein interaction sites (15,17).

The target structure is then scanned against a database of helix–turn–helix (HTH) templates taken from PDB structures known to be involved in DNA binding (18). Each template consists of the Cα coordinates of the HTH motif. Many false positives are returned by this search, but the majority can be filtered out if their solvent accessibility is below a certain cutoff.

The last in this batch of processes is a search for structural motifs called 'nests'. A 'nest' is an anion or cation binding site formed by three or more amino acids in the sequence whose

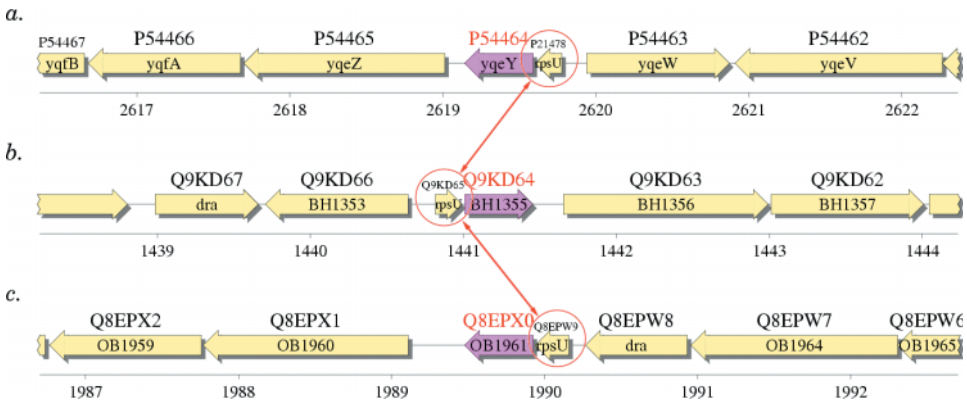

Figure 2. An extract from the genome neighbour analysis for a structure of unknown function. The structure is that of hypothetical protein YqeY from *Bacillus subtilis* (PDB code 1ng6). The Pfam annotation suggests that the protein might have a role in tRNA metabolism, but the genome analysis shown hints at the possibility of it being a ribosomal protein. The schematic diagrams show the neighbouring genes in (**a**) *B.subtilis*, (**b**) *Bacillus halodurans* and (**c**) *Oceanobacillus iheyensis*. Each gene is represented by an arrow indicating the length of the gene and its direction of transcription. The YqeY protein and its orthologues in the two other genomes are coloured pink, the sequence similarities being 75.7% to the *B.halodurans* protein and 69.4% to the *O.iheyensis* protein. In all three genomes, the gene of interest is immediately preceded by the rpsU gene, which corresponds to the 30S ribosomal protein S21. This pattern occurs in other bacterial genomes (data not shown), and suggests that the protein might be a ribosomal protein.

main-chain ψ–φ dihedral angles alternate between the right- and left-handed α and γ regions of the Ramachandran plot. Such motifs are found to be frequently associated with the functional sites of proteins (19,20).

3D template searches

The final batch of searches to which the target structure is subjected is the group of 3D template searches listed in the right-hand column of Figure 1. The templates used and the method of their detection and scoring are described in detail elsewhere (R. A. Laskowski, J. D. Watson and J. M. Thornton, manuscript submitted), but are summarized below.

We use four different types of templates: enzyme active sites, ligand-binding sites, DNA-binding sites and 'reverse' templates generated from the target structure itself. All templates consist of specific 3D conformations of between two and five amino acid residues. A fast 3D search program called Jess (21) is used to rapidly locate relative conformations of groups of residues in a given PDB file that closely match a specific template, reporting the r.m.s.d. between the matched and template residues. The comparison involves the side-chain atoms and, for the smaller residues, one or more of the main-chain atoms. For symmetrical side chains, the alternative conformations are also taken into account in the search.

The hits detected by Jess are then scored and ranked. The r.m.s.d. between the matched and template residues turns out not to be a particularly good measure for discriminating between true and false positives. A much better measure is the similarity between the local environments surrounding the matched residues in the target structure and the template's parent structure. This is computed by first pairing up identical residues in equivalent positions in the two proteins within a 10 Å sphere of the centre of the template match. Then, the paired residues are filtered to leave only those pairs that could have come from a sequence alignment of the two proteins in question (i.e. the residues in each pair are in the same relative sequential order in their respective sequences). Many different sets of pairings are possible, so each is scored and the highest scoring one is chosen. The score takes into account the number of paired residues in the supposed sequence alignment and the number of insertions that would be required in one or both of the sequences to arrive at this alignment.

Such a local similarity score is capable of identifying even quite distant homologues in cases where the residues involved in the functional site have been well conserved over evolutionary time, despite a marked divergence in the remainder of the proteins' sequences, and even their structures. So, for example, proteins having a sequence identity of 20–30% overall can have a much higher sequence identity, of say 40–50%, in the region of the functional site (R. A. Laskowski, J. D. Watson and J. M. Thornton, manuscript submitted).

Enzyme active sites. The first group of templates used in the template searches come from a manually curated database of the 3D conformations of enzyme active site residues. Each consists of 2–5 residues. These are the residues known from the literature to be catalytic, plus one or more additional residues whose 3D positions are highly conserved relative to the catalytic residues. The template database was originally called PROCAT and searched using a program called TESS (22), but has now been superseded by the Catalytic Site Atlas (23) and is searched by Jess (21). The database currently contains ~400 enzyme active site templates (http://www.ebi.ac.uk/thornton-srv/databases/CSA). Figure 3 shows how a match to one of these templates can provide particularly strong functional information.

Ligand-binding sites. The ligand-binding site templates identify the 3D conformations of residues that bind specific Het Groups in structures in the PDB. The PDB contains 5500 different Het Groups in complexes with proteins. For every type of Het Group, a dataset of non-homologous proteins binding that group are selected and used for automatically generating a set of three-residue ligand-binding templates. Each structure can generate one or more templates; the rules being that the residues forming the template must be interacting with the Het Group, that no residue in the triplet

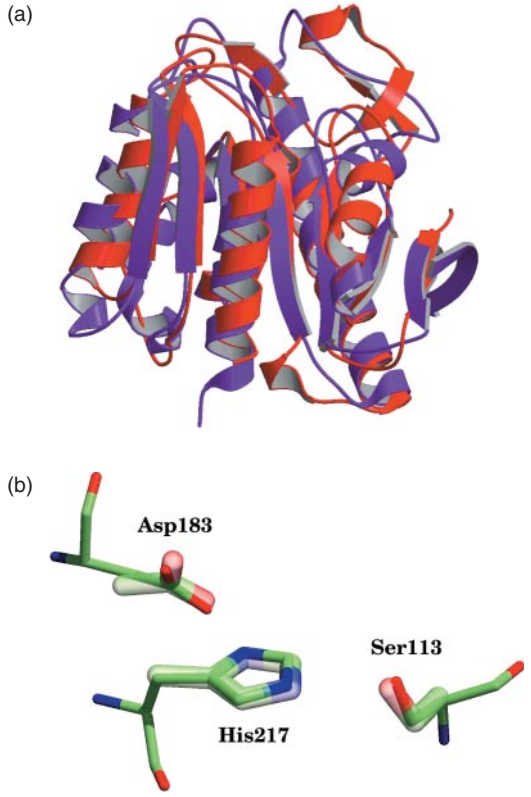

Figure 3. A simple case where structural information can provide strong confirmatory evidence for a hypothesized function. The example shown is of PDB entry 1ufo a hypothetical protein from *Thermus thermophilus* suspected of being a hydrolase. (**a**) The protein's fold matches a number of carboxylesterases, the match shown being to PDB structure 1aur (in red) which shares only a 24% sequence identity with 1ufo, but matches 11 SSEs and gives an r.m.s.d. of 1.83 Å on equivalent Cα coordinates when the two structures are superposed. (**b**) The protein returns a strong match to the enzyme active site template for the serine proteases. The Ser-His-Asp template is shown by the thick transparent bonds, while the matching side chains from the 1ufo structure are shown as thin solid bonds. The r.m.s.d. between the template and 1ufo atoms is 0.29 Å, which is well below the 1.2 Å cutoff for this template. Furthermore, when the 1ufo structure is superposed on the 1tah structure from which the template has been derived, there are 12 identical and 12 similar residues in equivalent 3D positions in the two structures within 10 Å of the template's centre. This high degree of local similarity in the region of the matched residues suggests that they could correspond to a Ser-His-Asp catalytic triad and that the hypothetical protein is a serine protease.

may be >5.0 Å from both the others, that no two templates from the same structure have more than two residues in common, and that none of the templates contains more than a single hydrophobic residue. The last of these restrictions aims at biasing the templates to contain mainly surface residues. As of February 2005, the database contained 13 057 ligand-binding templates.

DNA-binding sites. The steps for generating templates for the DNA-binding sites are identical to those for the ligand-binding templates. The data come from a non-homologous dataset of protein–DNA complexes. As of February 2005, the database contained 1200 DNA-binding templates.

'Reverse' templates. The fourth and final template method employed by ProFunc is the 'reverse' template method implemented in a program called SiteSeer. Rather than scan the target structure against a prepared database of templates the target itself is first broken up into a large set of several hundred templates. These are then scanned against a representative set of the structures in the PDB and any hits are scored and ranked as before. SiteSeer is able to find matches that the other template methods miss and, specifically, tends to match functionally important sites—as these are the ones most likely to have been preserved and hence give the highest local similarity scores. The list of representative structures scanned by SiteSeer is updated weekly using the Pisces server (24), obtaining a list of protein chains that are no more than 90% sequence identical and come from structures solved by X-ray crystallography at 3.0 Å resolution or better. The list contained 11 750 individual protein chains at February 2005.

ProFunc PROCESSING

Many of the processes that make up ProFunc are computationally demanding, particularly the Superfamily sequence search and the SiteSeer 'reverse' template search. Where possible the processes are run in parallel on the EBI's linux processor farms. A number of the searches are farmed off to other servers using SOAP (simple object access protocol) (25).

These include the BLAST sequence searches, sent to the EBI's BLAST server, and the fold search, sent to the EBI's SSM server. Use of SOAP has two main advantages: first, the task of maintaining the data and programs for a specific application is the responsibility of the server administrators and can be entrusted to them; and second, the computation is performed on someone else's processor(s), further distributing the overall processing load. The primary disadvantage is that one has no control over the external server used and therefore the pipeline is badly affected when the servers are down.

ProFunc OUTPUT

Depending on how busy the processor farm is, the ProFunc run on a single structure can take between half an hour and several hours. The server reports on progress and makes the results from the different processes available as soon as they are ready. Results are presented in summary form on the results page, with greater detail available on additional pages. Many of the analyses are supported by RasMol views of the specific matches in the target structure (e.g. location of template hits, superposition of matching folds, etc.).

SUMMARY

In bringing together a number of sequence and structure-based methods, the ProFunc server is a convenient tool for use in structural genomics. One submits a new structure to the server and, within a couple of hours, gets a number of complementary analyses relating to the protein's possible function. It is also likely to be useful for general analysis of newly solved structures as it can speedily identify sequence, structural and possibly functional relationship between the new structure and those already in the PDB. The server was developed as part of the structural genomics pipeline of the Midwest Center for Structural Genomics and has been running for over a year. It can be accessed at http://www.ebi.ac.uk/thornton-srv/databases/ProFunc.

A couple of improvements are planned for the near future. First, a summary of the hits from all the methods run on a given structure will be provided in terms of common functional terms. This should make it easier to see where the methods are consistently coming to the same, or similar, conclusions. The terms will be drawn not only from the protein names, as given in the UniProt, PDB, InterPro, etc. entries from which the hits have come, but also from any available functional annotations from these databases and from the Gene Ontology (26,27).

A second improvement will be to combine the ProFunc results with PDBsum analyses, generated specifically for the structure. The PDBsum database (28) already has a feature that enables upload of a structure for the generation of secure PDBsum pages and its analyses could greatly help in the study of any newly solved protein structure.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Martin Senger for his SOAP interface to the BLAST server and Dr Siamak Sobhany for access to his SSM SOAP server. This work was performed with funding from the National Institutes of Heath, grant number GM62414, the US DoE under contract W-31-109-Eng-38 and also as part of the BioSapiens Project which is funded by the European Commission within its FP6 Programme, under the thematic area 'Life sciences, genomics and biotechnology for health', contract number LHSG-CT-2003-503265. Funding to pay the Open Access publication charges for this article was provided by the former grant.

Conflict of interest statement. None declared.

REFERENCES

1. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
2. Berman,H.M. and Westbrook,J.D. (2004) The impact of structural genomics on the Protein Data Bank. *Am. J. Pharmacogenomics*, **4**, 247–252.
3. Watson,J.D., Laskowski,R.A. and Thornton,J.M. (2005) Predicting protein function from structure and structural data. *Curr. Opin. Struct. Biol.*, in press.
4. Laskowski,R.A., Watson,J.D. and Thornton,J.M. (2003) From protein structure to biochemical function? *J. Struct. Funct. Genomics*, **4**, 167–177.
5. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
6. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
7. Valdar,W.S.J. and Thornton,J.M. (2001) Conservation helps to identify biologically relevant crystal contacts. *J. Mol. Biol.*, **313**, 399–416.
8. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
9. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan: an integration platform for the signature-recognition methods in Inter Pro. *Bioinformatics*, **17**, 847–848.
10. Madera,M., Vogel,C., Kummerfeld,S.K., Chothia,C. and Gough,J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.*, **32**, D235–D239.
11. Krissinel,E. and Henrick,K. (2003) Protein structure comparison in 3D based on secondary structure matching (SSM) followed by Cα alignment, scored by a new structural similarity function. In Kungl,A.J. and Kungl,P.J. (eds), *Proceedings of the 5th International Conference on Molecular Structural Biology*, 3–7 September. Vienna, Austria, p. 88.
12. Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities and intermolecular interactions. *J. Mol. Graph.*, **13**, 323–330.
13. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374–376.
14. Lichtarge,O. and Sowa,M.E. (2002) Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.*, **12**, 21–27.
15. Madabushi,S., Yao,H., Marsh,M., Kristensen,D.M., Philippi,A., Sowa,M.E. and Lichtarge,O. (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.*, **316**, 139–154.
16. Glaser,F., Pupko,T., Paz,I., Bell,R.E., Bechor,D., Martz,E. and Ben-Tal,N. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.
17. Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.
18. Jones,S., Barker,J.A., Nobeli,I. and Thornton,J.M. (2003) Using structural motif templates to identify proteins with DNA binding function. *Nucleic Acids Res.*, **31**, 2811–2823.
19. Watson,J.D. and Milner-White,E.J. (2002) A novel main-chain anion-binding site in proteins: the nest. A particular combination of phi, psi values in successive residues gives rise to anion-binding sites that occur commonly and are found often at functionally important regions. *J. Mol. Biol.*, **315**, 171–182.
20. Watson,J.D. and Milner-White,E.J. (2002) The conformations of polypeptide chains where the main-chain parts of successive residues are enantiomeric. Their occurrence in cation and anion-binding regions of proteins. *J. Mol. Biol.*, **315**, 183–191.
21. Barker,J.A. and Thornton,J.M. (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics*, **19**, 1644–1649.
22. Wallace,A.C., Borkakoti,N. and Thornton,J.M. (1997) TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.*, **6**, 2308–2323.
23. Porter,C.T., Bartlett,G.J. and Thornton,J.M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.
24. Wang,G. and Dunbrack,R.L.,Jr (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.
25. Senger,M., Rice,P. and Oinn,T. (2003) Soaplab—a unified Sesame door to analysis tools. In Cox,S.J. (ed.), *Proceedings, UK e-Science, All Hands Meeting*, 2–4 September. Nottingham, UK, pp. 509–513.
26. The Gene Ontology Consortium (2000), Gene Ontology tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
27. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.
28. Laskowski,R.A., Chistyakov,V.V. and Thornton,J.M. (2005) PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res.*, **33**, D266–D268.