*doi:10.1093/nar/gki499*

# pdbFun: mass selection and fast comparison of annotated PDB residues

**Gabriele Ausiello*, Andreas Zanzoni, Daniele Peluso, Allegra Via and Manuela Helmer-Citterich**

Centre for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata,
Via della Ricerca Scientifica, 00133 Rome, Italy

Received February 14, 2005; Accepted April 28, 2005

## ABSTRACT

**pdbFun (http://pdbfun.uniroma2.it) is a web server for structural and functional analysis of proteins at the residue level. pdbFun gives fast access to the whole Protein Data Bank (PDB) organized as a database of annotated residues. The available data (features) range from solvent exposure to ligand binding ability, location in a protein cavity, secondary structure, residue type, sequence functional pattern, protein domain and catalytic activity. Users can select any residue subset (even including any number of PDB structures) by combining the available features. Selections can be used as probe and target in multiple structure comparison searches. For example a search could involve, as a query, all solvent-exposed, hydrophylic residues that are not in alpha-helices and are involved in nucleotide binding. Possible examples of targets are represented by another selection, a single structure or a dataset composed of many structures. The output is a list of aligned structural matches offered in tabular and also graphical format.**

## INTRODUCTION

Structural genomics projects (1) and the improvement of experimental techniques for structural analysis enrich the Protein Data Bank (PDB) (2) with structural data of very high quality and reliability. Nevertheless, few complete resources are available for analysing the connections between structural features and molecular functions that lie hidden in this huge amount of data. We identified some important characteristics that may be considered in the design and construction of a complete resource for establishing structure–function links: (i) presence of integrated data (number and type of different considered databases); (ii) level of the data integration detail (i.e. structure, domain, residue and atom); (iii) level of integration between data and computational tool(s) in the resource and (iv) wholeness (the quantity of data that can be analysed at the same time).

(i) Data integration can provide consistent advantages in the analysis of protein structures, as demonstrated and exemplified by PDBSUM (3), a database providing a vast amount of information on the PDB entries. At present, the huge MSD project (4), merging all main databases with the PDB, represents the best implementation of this concept.

(ii) Data integration can operate at different levels. Large volumes of data about protein structure and function are currently available in the biologically relevant databases. Such data can be integrated at the protein level. More effectively, for a focus on molecular function, they can be mapped onto protein residues. Data integration at the residue level is exemplified by the possibility of querying for solvent-exposed amino acids located in the alpha-helices of a protein structure. This feature has already been used in the SURFACE database (5) and has now been extended by MSDmine (unpublished resource, http://www.ebi.ac.uk/msd-srv/msdmine).

(iii) Integration between data and one or more computational methods is a fundamental task. Such a task is achieved in tools where simple or complex selections of the integrated data can be built and straightforwardly used as input to an embedded method (i.e. running a comparison program only on proteins sharing a specified function).

(iv) The last important property for a complete structural analysis tool is its being able to consider vast amounts of data at the same time, i.e. its wholeness or ability to work as a high-throughput resource. Queries can be formulated with more or even all the available data. A user may choose to focus on all proteins belonging to a specified SCOP class or to select all the tryptophan residues in the whole PDB catalytic sites.

In the perspective described here, we propose pdbFun as a fast and user-friendly integrated web server for structural analysis

*To whom correspondence should be addressed. Tel: +39 06 72594314; Fax: +39 06 72594314; Email: gabriele@cbm.bio.uniroma2.it of local similarities among proteins. pdbFun collects annotations derived from different databases (data integration), maps them onto single residues (good level of integration detail) and runs a local structural comparison algorithm on the selected residues (data/method integration). Queries and comparisons are allowed on any sets of annotations or residues, even including the entire PDB (wholeness).

## Overview

pdbFun is an integrated web tool for querying the PDB at the residue level and for local structural comparison. pdbFun integrates knowledge about single residues in protein structures from other databases or calculated with available instruments or instruments developed in-house for structural analysis. Each set of different annotations represents a feature. Typical features are secondary structure assignments or SMART domains (6), whose annotations are the H/T/E assignments or domain families, respectively, reported at the residue level. The user can build simple residue selections by including any number of annotations from a single feature, e.g. all residues belonging to any of three different SMART domains. The selections can be combined recursively to create more complex ones. The user is allowed to choose only the β-strand or turn residues of the previous three domains. Each selection can be manually refined by adding and removing single residues. Structural similarity can be searched between any pair of selections. All comparisons and queries are performed in real time with a fast program (Ausiello,G., Via,A. and Helmer-Citterich,M., manuscript submitted) running on the web server.

## Features

The different features currently available are shown on, and can be accessed from, the homepage. The user can start creating one residue selection by choosing any one of the following (Figure 1):

(i) *Structures*. All residues belonging to one or many PDB structures can be selected, up to and including the whole database.

(ii) *Chains*. All residues belonging to one or more chains can be selected. Lists of non-redundant PDB chains are available here as pre-calculated selections.

(iii) *Surfaces*. Residues can be selected according to their solvent-exposed or buried status given by the NACCESS program (7).

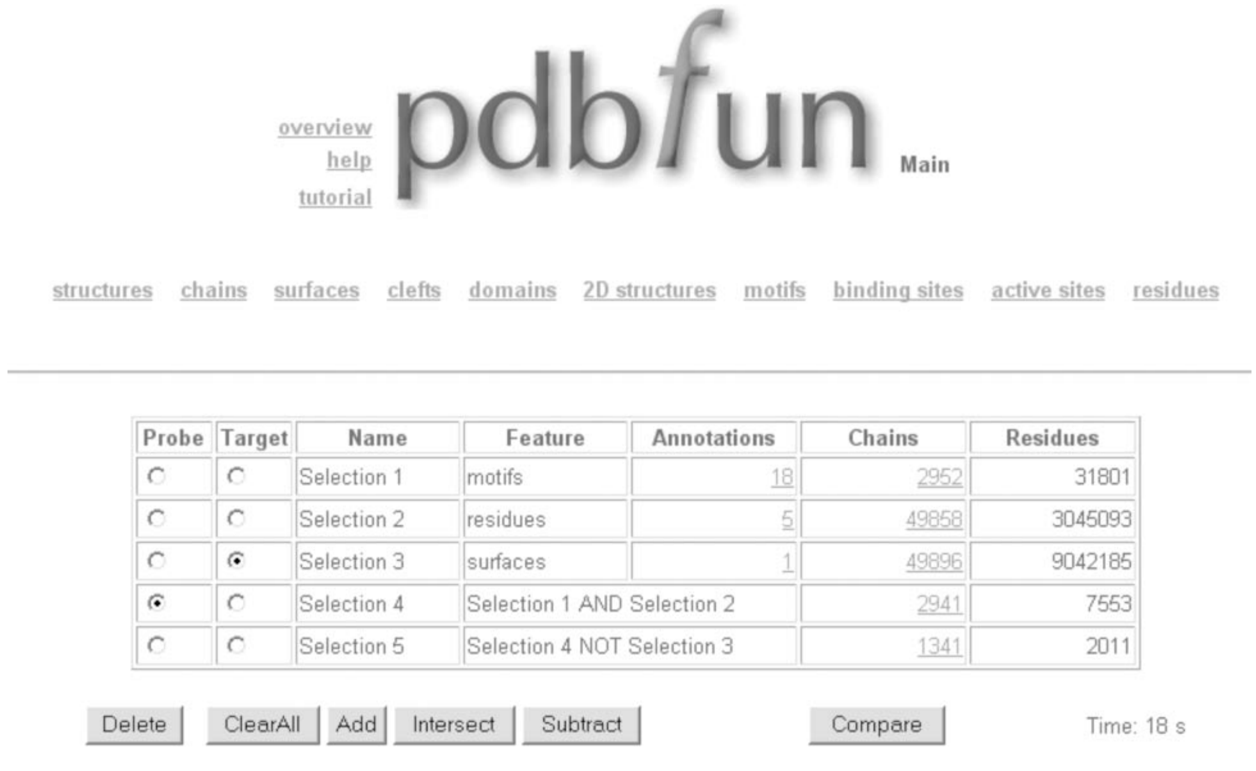
(iv) *Clefts*. The SURFNET program (8) is used to assign surface residues to protein cavities. Cavities are sorted by size (number 1 refers to the biggest).

(v) *Domains*. Residues belonging to domains are annotated here using HMMER (9) on the SMART database.

(vi) *Two-dimensional structures*. Each residue is associated with the secondary structure assignment provided by the dssp (10) program. (E: extended strand; H: alpha-helix; T: hydrogen bonded turn, etc.).

(vii) *Motifs*. PROSITE patterns (11) as found on the sequences of the PDB chains.

(viii) *Binding sites*. Users can select residues whose distance is <3.5 Å from any ligand molecule present in the PDB. Choosing ATP or ADP selects all residues found at a



| Probe | Target | Name | Feature | Annotations | Chains | Residues |
|---|---|---|---|---|---|---|
| ○ | ○ | Selection 1 | motifs | 18 | 2952 | 31801 |
| ○ | ○ | Selection 2 | residues | 5 | 49858 | 3045093 |
| ○ | ● | Selection 3 | surfaces | 1 | 49896 | 9042185 |
| ● | ○ | Selection 4 | Selection 1 AND Selection 2 | | 2941 | 7553 |
| ○ | ○ | Selection 5 | Selection 4 NOT Selection 3 | | 1341 | 2011 |

Time: 18 s

**Figure 1.** A Selection table is shown. The user has created five selections: Selection 1, all PROSITE residues with the ATP keyword in the pattern description (using the motifs feature); Selection 2, all charged residues in the PDB (D, E, H, K and R in the residues feature); Selection 3, all exposed residues (surface feature); Selection 4, all charged residues in the selected motifs (Selection 1 INTERSECT Selection 2); Selection 5, all charged residues in the selected motifs that are not solvent-exposed (Selection 4 SUBTRACT Selection 3). The estimated time for comparing the first chain (see text) of Selection 5 as query and Selection 3 as target is 18 s.

distance closer than the defined threshold from the ATP or ADP nucleotides.

(ix) *Active sites*. Active site residues in a set of enzyme structures obtained from the CatRes database (12).

(x) *Residues*. The 20 residue types [from A (alanine) to W (tryptophan)]. This feature helps the user to concentrate only on some kinds of residues, while ignoring all the others (i.e. all charged residues or aromatic residues).

### Annotations

By selecting a feature from the pdbFun main page, the user accesses the annotation page where single annotations of that particular feature can be chosen to create a simple selection of residues. The total number of selected residues corresponds to the sum of all the residues selected by a single annotation. We describe in detail the Motifs feature page.

In the Motifs page, all PROSITE patterns are listed and represent the annotations. Fields duplicated locally are the pattern 'id', 'name' and 'short description'. In addition, a 'residues' field indicates the number of annotated residues in the whole PDB. A 'chains' field indicates the number of chains containing at least one of the annotated residues. In order to facilitate searching, the annotations are organized in pages and can be sorted by any field.

Annotations (i.e. specific PROSITE motifs) can be added to the current selection in various ways: manually (using check-boxes), by text search (only the selected field will be searched) or by uploading a user flat file containing a list of PROSITE codes.

All the features available in pdbFun share identical organization. New features can therefore be added and annotations handled without the need to modify the code.

Let us take as an example how to select all PDB residues matching any of the PROSITE motifs involving ATP.

(i) In the Motifs page, sort the annotations by the 'description' field by clicking on the column header.

(ii) Type 'ATP' in the search box (the search will be automatically conducted on the sorted field) and press the search button.

(iii) All the 18 PROSITE motifs containing 'ATP' in the description are selected, and the user can go back to the main page and find the selection described as a row on the pdbFun main page.

### Simple selections

Whenever a selection is made, pdbFun stores it as a row in a Selection table that can be visualized by going back to the main page. Each selection is identified by a unique name, by a type (the feature used to generate it), by the number of annotations selected in the feature and by the total number of chains and residues in the PDB that have been selected. New selections can be created by choosing one of the features available in the upper part of the screen. Existing selections can be accessed and modified via the 'annotations' field.

For example, see Figure 1. The selection created in the previous example now appears in the Selection table as 'Selection 1'. The 'feature type' field is 'motifs'. The number of annotations selected is 18 (the 18 PROSITE patterns whose description contains the ATP word). Such patterns have been found on 2952 different PDB chains and comprise a total of 31 801 residues.

### Combining selections

All selections can be combined using the AND, OR and NOT boolean operators. The result is a new selection containing a combination of their residues. The two selections to combine are chosen with the 'probe' and 'target' radio buttons. Applying the 'Intersect' (AND) on Selections 1 and 2 (see Figure 1) creates a new selection including only the common residues (e.g. the PDB proline residues that are found in alpha-helices), whereas using 'Add' (OR) the two selections will be merged (e.g. all residues that are in a big surface cleft 'or' belong to some active site). The 'Subtract' (NOT) is also a binary operator and needs to be understood as an 'AND' between the probe and the complement of the target (e.g. all the charged residues which are 'not' exposed).

Each selection created can be, recursively, the object of a new combination.

The 'residues' and 'chains' columns of the Selection table contain useful statistical information on the PDB residues' composition. Questions such as 'How many charged residues are buried in the whole PDB, or in a certain type of domain?' can be answered in a fraction of a second.

### Structural comparison

Selections can be chosen as probe and target of a structural comparison procedure to find local similarities in residues' spatial arrangements. The selected residues in each chain of the probe are searched against the selected residues in each chain of the target. The comparison algorithm is guaranteed to find the largest subset of matching residues between two structures. The matching condition is an RMSD (root mean square difference) <0.7 Å and a residue similarity >1.3 according to the Dayhoff substitution matrix. The algorithm is exhaustive, fast and sequence and fold independent.

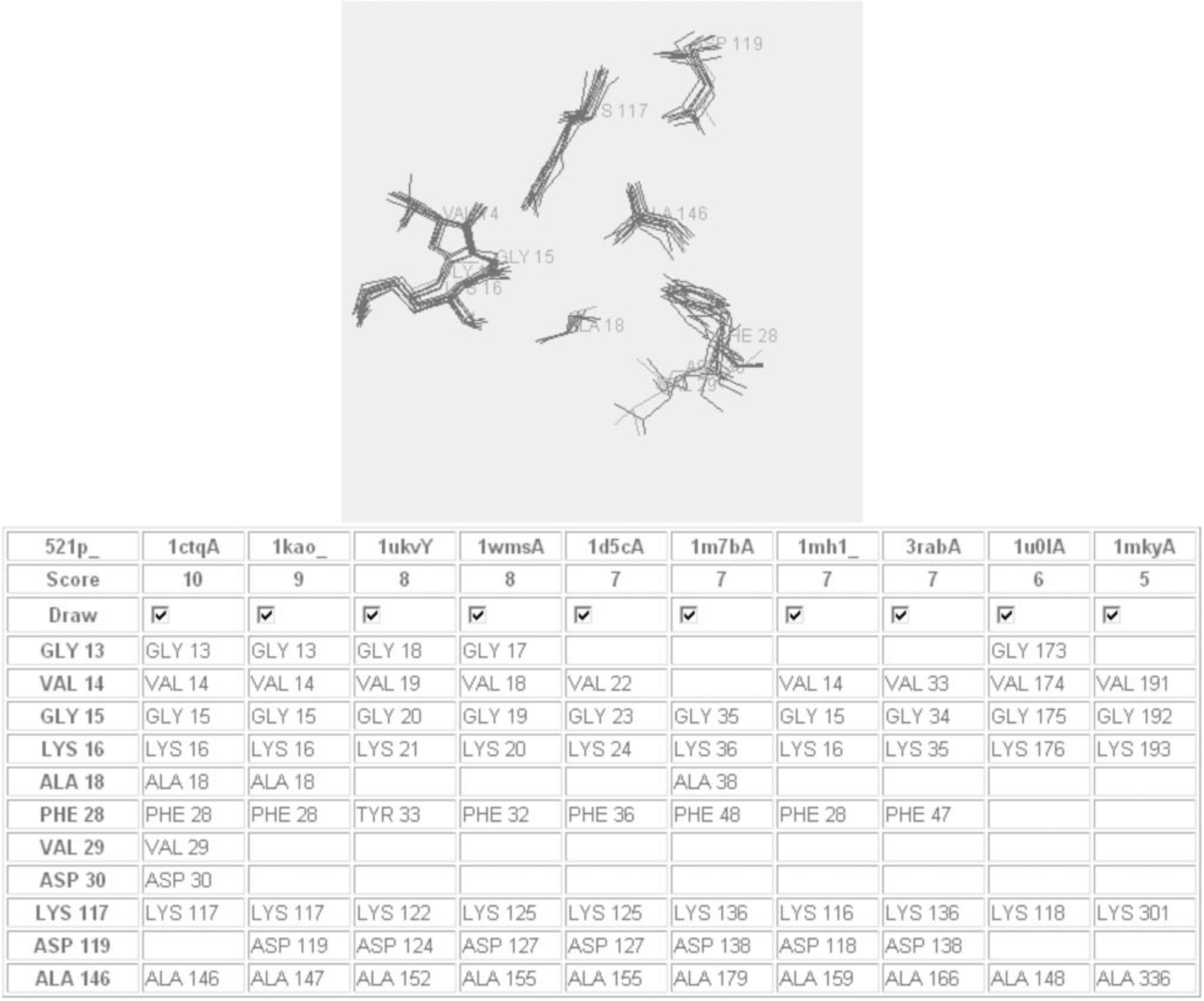
All the probe (but not the target) residues must belong to a single PDB chain (if the probe is a multi-chain selection, only the first chain will be compared by default). Comparisons stop when a match is found comprising at least 10 residues. As soon as a new probe or target is chosen, an estimate of the comparison execution time is given at the bottom right of the screen.

### Comparison results

When a comparison is started, the user is redirected to the Results page. Here new matches are immediately displayed as they are calculated. Matches are sorted by decreasing score and are displayed in pages. The probe chain matching residues are listed in the first column of the Results table. Each target chain is shown in a different column, together with the match length. Target residues are listed in the same rows as the probe residues to which they are structurally aligned (see Figure 2). Columns can be selected for a graphical view of the match in single or multiple alignment using a Java applet. A text file containing the results of the comparison is available for downloading.

### Manual selections

pdbFun allows the user to perform a manual selection of residues on a single PDB chain, according to his/her interest



| 521p_ | 1ctqA | 1kao_ | 1ukvY | 1wmsA | 1d5cA | 1m7bA | 1mh1_ | 3rabA | 1u0lA | 1mkyA |
|---|---|---|---|---|---|---|---|---|---|---|
| Score | 10 | 9 | 8 | 8 | 7 | 7 | 7 | 7 | 6 | 5 |
| Draw | ☑ | ☑ | ☑ | ☑ | ☑ | ☑ | ☑ | ☑ | ☑ | ☑ |
| GLY 13 | GLY 13 | GLY 13 | GLY 18 | GLY 17 | | | | | GLY 173 | |
| VAL 14 | VAL 14 | VAL 14 | VAL 19 | VAL 18 | VAL 22 | | VAL 14 | VAL 33 | VAL 174 | VAL 191 |
| GLY 15 | GLY 15 | GLY 15 | GLY 20 | GLY 19 | GLY 23 | GLY 35 | GLY 15 | GLY 34 | GLY 175 | GLY 192 |
| LYS 16 | LYS 16 | LYS 16 | LYS 21 | LYS 20 | LYS 24 | LYS 36 | LYS 16 | LYS 35 | LYS 176 | LYS 193 |
| ALA 18 | ALA 18 | ALA 18 | | | | ALA 38 | | | | |
| PHE 28 | PHE 28 | PHE 28 | TYR 33 | PHE 32 | PHE 36 | PHE 48 | PHE 28 | PHE 47 | | |
| VAL 29 | VAL 29 | | | | | | | | | |
| ASP 30 | ASP 30 | | | | | | | | | |
| LYS 117 | LYS 117 | LYS 117 | LYS 122 | LYS 125 | LYS 125 | LYS 136 | LYS 116 | LYS 136 | LYS 118 | LYS 301 |
| ASP 119 | | ASP 119 | ASP 124 | ASP 127 | ASP 127 | ASP 138 | ASP 118 | ASP 138 | | |
| ALA 146 | ALA 146 | ALA 147 | ALA 152 | ALA 155 | ALA 155 | ALA 179 | ALA 159 | ALA 166 | ALA 148 | ALA 336 |

**Figure 2.** The first Results page of a comparison. A manual selection of 5p21 (ras protein) residues involved in GTP binding was compared with the ∼5500 chains of a non-redundant PDB (50%). The output is shown in tabular and also graphic format. In the first column of the table, the matching residues of the query PDB chain are reported; in the adjacent columns, the other PDB chains follow, and the residues aligned in three dimensions appear in the same rows. The matched PDB chains are reported in the first row; the number of matched residues in the second one. Matching residues are also displayed upon selection (pressing on the 'draw' button) with a Java applet.

or personal knowledge (and not only by using the features calculated or extracted from pre-existing databases). Through the 'chains' field in the Selection table, the user accesses a page where he/she can choose the chain to work with manually.

All the residues in the chain of interest will appear as a list, together with the available annotations. Sets of single residues can be chosen. A simple Java applet helps the user in selections. This selection appears in the Selection table as 'manual selection'.

## Non-redundant PDB sets

Non-redundant datasets of chains obtained from the PDB (2) at different (90, 70, 50 and 30%) redundancies are available and can be used to generate non-redundant selections of chains or as target datasets. These sets can be selected from the Chains feature page and modified manually or left as they are.

## Implementation notes

In order to achieve high speed and a high level of interactivity, all residue data are stored in the server memory. A single C program executes both fast queries and structural comparisons, and a relational database is used only for the storage of the feature annotations list and for web users management. All selections can be run in a fraction of a second. Comparison times range from fractions of a second to minutes. No time limit is given to users (but a newly submitted job stops the running one). Web pages have been tested on the main browsers for the Windows and Linux platforms. Mac users should utilize Safari ⩾1.2.

## Future directions

Major future developments involve the addition of new features. Features in preparation are residue conservation derived by HSSP (13), presence in structural fold derived by CATH (14), user-defined sequence regular expressions and proximity of residues. Finally, to further improve the quality of integration among different data sources, part of the MSD data collection could be used.

Upload of user structures will be made possible and statistical significance of the matches introduced.

## ACKNOWLEDGEMENTS

We thank Federico Fratticci for his useful contribution. This work was supported by Telethon project GGP04273, FIRB and Genefun. Funding to pay the Open Access publication charges for this article was provided by Telethon project GGP04273.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Skolnick,J., Fetrow,J.S. and Kolinski,A. (2000) Structural genomics and its importance for gene function analysis. *Nat. Biotechnol.*, **18**, 283–287.
2. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33**, D233–D237.
3. 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.
4. Velankar,S., McNeil,P., Mittard-Runte1,V., Suarez,A., Barrell,D., Apweiler,R. and Henrick,K. (2005) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.*, **33**, D262–D265.
5. Ferrè,F., Ausiello,G., Zanzoni,A. and Helmer-Citterich,M. (2004) SURFACE: a database of protein surface regions for functional annotation. *Nucleic Acids Res.*, **32**, 240–244.
6. Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P. and Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.*, **32**, D142–D144.
7. Hubbard,S.J. and Thornton,J.M. (1993) NACCESS Computer Program. Department of Biochemistry and Molecular Biology, University College London.
8. Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities and intermolecular interactions. *J. Mol. Graph.*, **13**, 323–330.
9. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
10. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
11. Hulo,N., Sigrist,C.J.A., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32**, D134–D137.
12. Bartlett,G.J., Porter,C.T., Borkakoti,N. and Thornton,J.M. (2002) Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.*, **324**, 105–121.
13. Sander,C. and Schneider,R. (1991) Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
14. Pearl,F.M.G., Lee,D., Bray,J.E., Sillitoe,I., Todd,A.E., Harrison,A.P., Thornton,J.M. and Orengo,C.A. (2000) Assigning genomic sequences to CATH. *Nucleic Acids Res.*, **28**, 277–282.