Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W549–W554***
DOI: 10.1093/nar/gkh439

PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins

Vladimir A. Ivanisenko*, Sergey S. Pintus, Dmitry A. Grigorovich and Nickolay A. Kolchanov

Institute of Cytology and Genetics SBRAS, Lavrentyev Avenue 10, Novosibirsk 630090, Russia

Received February 14, 2004; Revised and Accepted April 16, 2004

ABSTRACT

PDBSiteScan is a web-accessible program designed for searching three-dimensional (3D) protein fragments similar in structure to known active, binding and posttranslational modification sites. A collection of known sites we designated as PDBSite was set up by automated processing of the PDB database using the data on site localization in the SITE field. Additionally, protein–protein interaction sites were generated by analysis of atom coordinates in heterocomplexes. The total number of collected sites was more than 8100; they were assigned to more than 80 functional groups. PDBSiteScan provides automated search of the 3D protein fragments whose maximum distance mismatch (MDM) between N, Cα and C atoms in a fragment and a functional site is not larger than the MDM threshold defined by the user. PDBSiteScan requires perfect matching of amino acids. PDBSiteScan enables recognition of functional sites in tertiary structures of proteins and allows proteins with functional information to be annotated. The program PDBSiteScan is available at http://wwwmgs.bionet.nsc.ru/mgs/systems/fastprot/pdbsitescan.html.

INTRODUCTION

To gain a better understanding of protein function, more has to be known about sites that might be active, binding or posttranslationally modified. Various methods have been developed for recognizing functional sites in protein primary structure (1), using combined data on multiple sequence alignment and on three-dimensional (3D) structure (2–4), and searching sites in 3D structure (5–10). Common to the methods was the involvement of the conserved properties of sites either at the level of amino acid sequences, or site geometry, or site physicochemical properties. All these methods required the inclusion of a large number of sites of known function in a learning set. Exhaustive search of functional sites has been based on the data on protein–ligand interactions. A good example is the RELIBASE database system (11).

With this in mind, we developed the PDBSiteScan program, which handles and searches active sites, binding sites and posttranslational modification sites in 3D protein structure by pairwise structural comparison of a protein with each representative of these sites from the PDBSite collection. PDBSite was developed using the data of the SITE field from the Protein Data Bank [PDB, (12)] in addition to analysis of atom coordinates in heterocomplexes for protein–protein interaction sites.

The results of the PDBSiteScan search were identification of all the 3D protein fragments similar in tertiary structure and amino acids at least to one of the sites. An advantage of PDBSiteScan is that it can be applied to site recognition regardless of the number of sites of known function. We used catalytic triads as an example, illustrating this approach applied with good accuracy to recognition of large groups of functional sites. Recognition of the potential Na-binding site in human DJ-1 protein demonstrated that PDBSiteScan is also efficient for small groups of functional sites.

MATERIALS AND METHODS

A collection PDBSite

A collection of functional sites, PDBSite, was generated by automated processing of the PDB fields, HEADER, TITLE, KEYWDS, REMARK 800, SITE and ATOM. The grammatical parsing programs were developed to process the PDB entries. When a PDB entry contained data for several sites, a separate entry was generated for each of the sites in PDBSite. The HEADER, TITLE and KEYWDS fields were used to generate the supplement data for proteins from which the sites were extracted. The description of the function of a site was extracted from the REMARK 800 field. A program was developed for automated functional classification of sites

*To whom correspondence should be addressed. Tel: +7 383 233 2971; Fax: +7 383 233 1278; Email: salix@bionet.nsc.ru

Table 1. A functional grouping of sites with an indication of the number of sites for each group used as templates for site recognition

Active sites	
Active sites	1389
Posttranslational modification sites	
Acetylation	5
Phosphorylation	23
Lipoylation	2
Myristylation	5
Cleavage	5
Glycosylation	13
Metal binding sites	
Cadmium	6
Calcium	431
Cobalt	1
Copper	32
Iron	60
Metal	49
Molybdenum	2
Manganese	55
Magnesium	73
Mercury	3
Nickel	23
Potassium	
Gallium	1
Sodium	6
Vanadium	6
Ytterbium	3
Zinc	259
Inorganic non-metal binding sites	
Carbohydrate	29
CO	17
Phosphate	54
Sulfur	247
Xenon	1
Organic ligand binding sites	
NAD	46
Ascorbate	5
ATP	113
Oxalate	2
Peptide	1
Xylan	3
HEC	150
Uracil	6
Hapten	1
Benzamidine	1
Benzhydrozamic	6
Butyramide	2
Citrate	2
Cyclosporin	22
Diaminopimelate	1
DTTP	3
FAD	66
Fluconazole	1
FMN	66
Formycin	2
Maltose	15
NADPH	44
NADP	26
Pterin	1
Purine	4
Purvalanol	1
Pyridoxal	4
Pyrimidine	4
Pyruvate	1
Saccharopine	8
Staurosporine	2
Succinate	1
Sugar	19
Zanamivir	2
Lactose	1
Ornithine	4
Isopropylmalate	1
Trisaccharide	1
Glycerol	1
Tris	1
Glutathione	16
Thymidine	3
Glucose	12
Thiamin	2
Ganglioside	5
Tetracycline	1
Lipid	4
Molybdopterin	16
HEME	1
Nucleotide	5
Pharmaceutical drug binding sites	
Cyclosporin	22
Zanamivir	2
Formycin	2
Staurosporine	2
Protein binding sites	
Actin	2
GTPase	12
Heparin	10
Cyclophilin	2
Protein–protein interaction sites extracted from heterocomplexes	1002
Miscellaneous sites	
Miscellaneous sites	3588

on the basis of their textual description; its current version is able to classify functional sites of about 80 types. All the unclassified sites were united into the Miscellaneous group, each site containing a description of the function that allows the user to recognize the functional class of a site.

In addition to the sites extracted from PDB using the SITE field, the protein–protein interaction sites identified from the calculated contact residues in heterocomplexes are collected in PDBSite. Only the PDB structures containing the word COMPLEX in the HEADER field were processed. The interaction sites were defined only for the contacts between the different proteins, while those between the subunits of the same protein were disregarded. In the current version, only the coordinates directly given in the PDB file were used. The proteins that required space group symmetry involving either rotation or translation were omitted from consideration. A residue was accepted as contact if it had at least three atoms whose distance from any atoms of the partner chain was <5 Å (13). All these sites were assigned to a group of the protein–protein interaction sites. Table 1 presents the results of the functional classification of the sites.

Functional site recognition

Each site from the PDBSite collection served as a template for functional site recognition; the introduced strategy relied on a search of fragments in the tertiary structure of proteins

structurally similar to a site template and having the same amino acids as the site template.

The CE algorithm concept (14) underlay the one we implemented here in the structural alignment of protein and site template. Three atoms N, Cα and C, which defined the orientation of residues in space, were under consideration (15). All the possible triads of residues in distinct regions of protein 3D structure were examined as candidates for the core of the search site. Then, the core of a site was subject to incremental combinatorial extension until the number of residues became equal to that of residues in the site template. Decision making, whether or not to add a residue to the core, was based on the estimated structural similarity and match between the residue types.

To estimate structure similarity, the structure of a template and a protein fragment was superimposed by Kabsch's method (16). The maximum distance mismatch (MDM) was estimated as

$$\mathrm{MDM} = \max\left\{\sqrt{\left(x_i^s - x_i^p\right)^2 + \left(y_i^s - y_i^p\right)^2 + \left(z_i^s - z_i^p\right)^2}\right\}_{i=1}^{n} \quad \mathbf{1}$$

where x_i^s, y_i^s, z_i^s are the coordinates of the *i*-th atom for the site template, x_i^p, y_i^p, z_i^p are the coordinates of the *i*-th atom for the protein fragment, *n* denotes the number of atoms for the corresponding list of residues. The order of atoms for each residue of the site template and protein was defined as N, Cα and C. In this way, only atoms of the same type were superimposed.

SOFTWARE ACCESS

The PDBSiteScan web page is available at http://wwwmgs.bionet.nsc.ru/mgs/systems/fastprot/pdbsitescan.html (Figure 1). The input PDBSiteScan are a PDB file with the 3D structure of the analyzed protein and the threshold value of the MDM. The user is also able to specify the functional group of sites to be searched. The output contains information on the superimposition of each site–protein fragment pair, with the structural differences between each member of the pair not exceeding a predefined threshold MDM value (Figure 2). It also includes a unique PDBSite identifier of the site, the protein PDB ID from which the site was extracted, site description, the MDM and root mean square deviation (RMSD) values for the identified protein fragment; structural alignment data. Each result is linked to the complete information on the site in PDBSite.

PDBSiteScan provides structure site-protein alignment as a PDB file. This allows visualization of the structure alignment by popular softwares, such as Chime and RasMol.

TESTING PDBSITESCAN

Recognition of catalytic sites in the hydrolase family was provided as an example for testing PDBSiteScan. Two samples of non-homologous proteins (an identity of 35% or less) of 23 proteins each were set up. The positive sample was composed of hydrolases whose catalytic sites are known

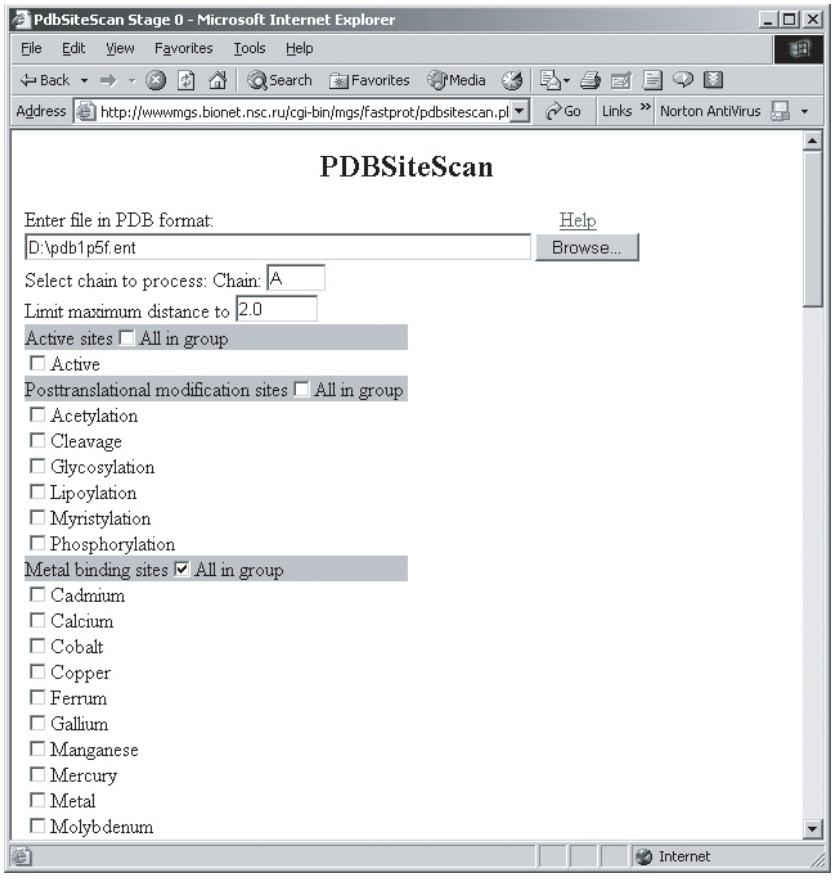

Figure 1. PDBSiteScan web page.

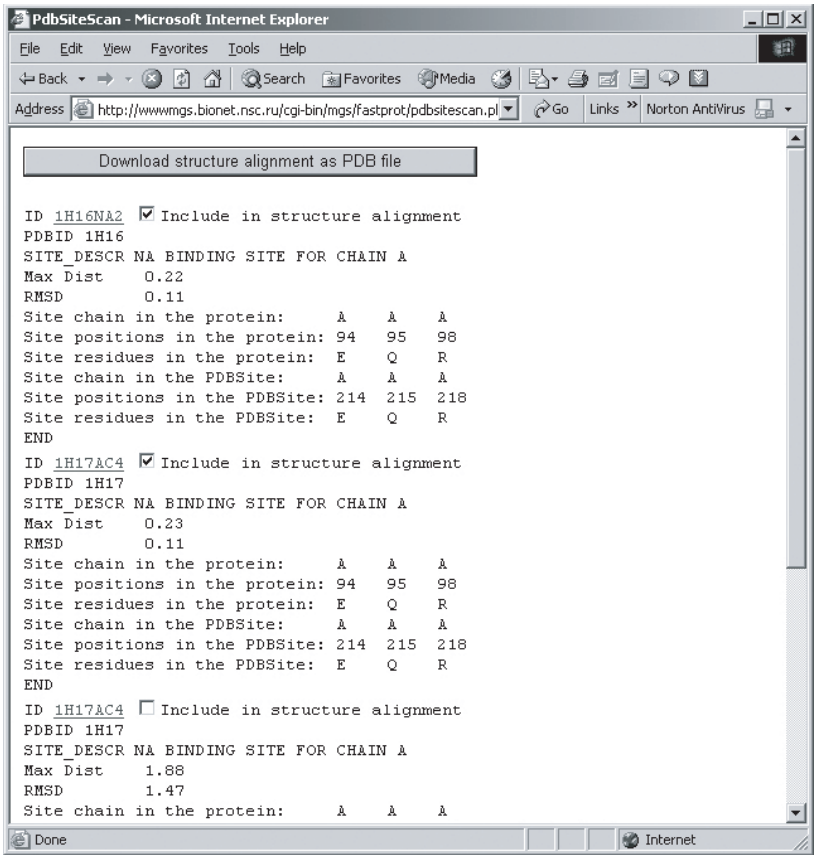

Figure 2. Screen capture of the PDBSiteScan output returned to the user.

Table 2. Catalytic sites of proteins of the hydrolase family used for testing PDBSiteScan

PDB ID	Catalytic site
1A2Z	80E, 143C, 167H
1AUO	114S, 168D, 199H
1B2L	138S, 151Y, 155K
1B6G	124D, 260D, 289H
1BDB	142S, 155Y, 159K
1BIF	256H, 325E, 390H
1BIO	57H, 102D, 195S
1BN7	117D, 141E, 283H
1BRT	98S, 228D, 257H
1BS9	90S, 175D, 187H
1CUJ	120C, 175D, 188H
1SVP	141H, 163D, 215A
1CV8	24C, 120H, 141N
1CVL	87S, 263D, 285H
1E6U	107S, 136Y, 140K
1E6W	155S, 168Y, 172K
1EA5	200S, 327E, 440H
1ELV	460H, 514D, 617S
1H2W	554S, 641D, 680H
1H4W	57H, 102D, 195S
1HAZ	57H, 102D, 195S
1QJ1	57H, 102D, 195S
1QJ4	80S, 207D, 235H

(Table 2), while the second negative was comprised of proteins not containing catalytic centers. PISCES (17) was used to set up the samples of non-homologous proteins. Only proteins of known 3D structure were included in analysis.

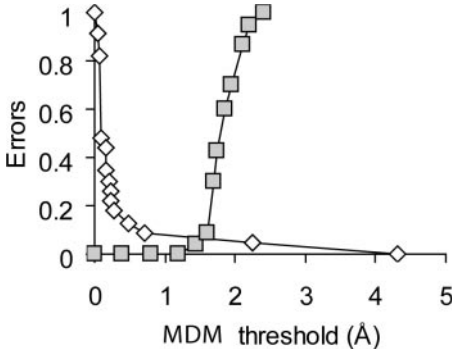

Figure 3. Graph showing the relation of type I (open diamonds) and type II errors (closed squares) for the catalytic site recognition in a hydrolase set to the threshold MDM values.

Using these samples, type I and type II errors were estimated for catalytic site recognition in proteins at different threshold values of the MDM. The total number of catalytic sites of hydrolase family in PDBSite was 297. Recognition was positive, when among the PDBSiteScan predicted sites at least one catalytic hydrolase site matched perfectly with the real site. The catalytic site of protein used as a template in the PDBSite collection was disregarded. Figure 3 is a graphical representation of the relation of type I and type II errors for the catalytic site recognition to the threshold MDM values.

Type I error was expressed as the portion of unidentified catalytic sites at a given value of the MDM threshold:

$$E_1(\mathrm{MDM}) = \frac{n(\mathrm{MDM})}{N_p} \quad \mathbf{2}$$

where n(MDM) is the number of unidentified catalytic sites in the positive sample, N_p = 23. Type II error was calculated using the formula

$$E_2(\mathrm{MDM}) = \frac{k(\mathrm{MDM})}{N_n} \quad \mathbf{3}$$

where k(MDM) is the number of proteins having at least one false catalytic site in the negative sample N_n.

The catalytic sites are recognized with good accuracy (Figure 3). Thus, at a threshold value for the MDM a little less than 1 Å, the number of positive identified sites exceeds 80% and that of the false is 0.

As seen in Figure 3, the optimum threshold values for the MDM are in the 1.0–1.5 Å range. The requirements for structural similarities between the potential and real sites became less stringent as these values decreased. Thus, at high threshold values for the MDM, the potential sites greatly differing in structure from the real were permissible among the predicted sites; the consequence was overprediction of the active sites. At lower threshold values for the MDM, the potential sites, even those structurally similar to the real, could be discarded by PDBSiteScan; the consequence was underprediction of the active sites. The optimum threshold MDM values of the other functional sites (Table 1) vary in the narrow 1.0–2.5 Å range, as optimized by trial and error.

PROTEINS OF UNKNOWN FUNCTION

Human DJ-1 was analyzed to illustrate the application of PDBSiteScan to the annotation of functions for proteins of interest. DJ-1 is a protein involved in multiple physiological processes, including cancer, Parkinson's disease and male fertility (18,19). Furthermore, DJ-1 was identified as a regulatory subunit (RS) of an RNA-binding protein (RBP) complex and inhibits the RNA-binding activity of RBP (20). It remains unclear how DJ-1 functions in the apparently different systems.

A potential Na-binding site containing positions E94, Q95, R98 (MDM = 0.22 Å) was identified in the structure of DJ-1/RS [PDB ID 1p5f (Figure 4)] using PDBSiteScan. Clearly, the potential site differs by the positions of the side chain atoms that are able to freely rotate around their covalent bonds.

The widespread effects of specific monovalent cation in biological macromolecules are highly relevant to physiological processes (21,22). Na^+ or K^+ is required for optimum catalysis of a number of enzymes; the cations act as allosteric effectors that alter the structure of the entire protein, or as cofactors binding to specific substrates (23). One must not neglect data indicating that despite their scarcity, enzymes activated by Na^+ participate in key regulatory interactions (24). Another thought-provoking example may be the triggering of molecular events by Na^+-specific effects resulting in functions of physiological importance: thrombin is allosterically switched from an anticoagulant—procoagulant factor (25) as a result of binding of Na^+ to a single site (26,27).

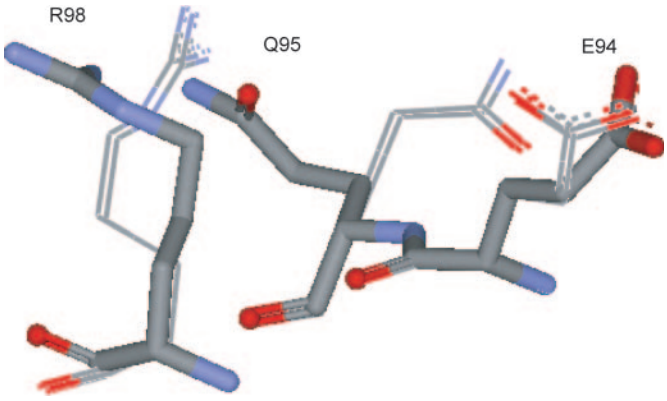

Figure 4. The results of a PDBSiteScan search demonstrate that the residues of a potential Na-binding site (stick model) of the DJ-1 protein superpose well with the real Na-binding sites (line model) of the PDBSite collection. The view of the sites was obtained using ViewerLite.

Taken together, this allows us to assume that sodium ion might perhaps have a role in the regulation of the RNA-binding complex. To our knowledge, the involvement of sodium ions in the regulation of the function of DJ-1 has not been suggested in the literature. Further tests of the Na-binding ability of DJ-1 using molecular dynamic modeling, e.g., are needed.

FUTURE DEVELOPMENTS

It is intended to obtain a more complete classification of protein functional sites. Other planned developments include programs for automated classification of sites on the basis of their textual description and structural comparison with the previously classified sites. A new class of protein–protein interaction will be derived from data on the contacts between the subunits in the polysubunit proteins that are the equivalents of biological units. In addition, it is intended to add an option for the ligand-binding sites in PDBSiteScan. The option will allow structural alignments and atom–ligand coordinates to be visualized simultaneously. These coordinates may be the starting points for molecular dynamics simulation or docking. It is proposed to increase the accuracy of site prediction by taking the site's environment into consideration.

ACKNOWLEDGEMENTS

We thank Anna Fadeeva for translating the manuscript from Russian into English. The work was partly supported by the Russian Foundation for Basic Research (grants 01-07-90376 and 03-07-96833-p2003); the Siberian Branch of the Russian Academy of Sciences (Integration Project No. 65); the U.S. Civilian Research & Development Fund for the Independent States of the former Soviet Union (CRDF); the Basic Research and Higher Education (BRHE) program NO-008-X1; MCB RAS (No. 10.4); the Siberian Branch of the Russian Academy of Sciences (Integration Project No. 119); and Russian Ministry of Industry, Science and Technologies (grant No. 43.073.1.1.1501).

REFERENCES

1. Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J.A., Hofmann,K. and Bairoch,A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
2. Fetrow,J.S. and Skolnick,J. (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.*, **281**, 949–68.
3. Todd,A.E., Orengo,C.A. and Thornton,J.M. (1999) DOMPLOT: a program to generate schematic diagrams of the structural domain organization within proteins, annotated by ligand contacts. *Protein Eng.*, **12**, 375–379.
4. Holm,L. and Sander,C. (1997) An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. *Proteins*, **28**, 72–82.
5. Liang,M.P., Banatao,D.R., Klein,T.E., Brutlag,D.L. and Altman,R.B. (2003) WebFEATURE: an interactive web tool for identifying and visualizing functional sites on macromolecular structures. *Nucleic Acids Res.*, **31**, 3324–3327.
6. 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.
7. Gutteridge,A., Bartlett,G.J. and Thornton,J.M. (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J. Mol. Biol.*, **330**, 719–734.
8. 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.
9. Artymiuk,P.J., Poirrette,A.R., Grindley,H.M., Rice,D.W. and Willett,P. (1994) A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J. Mol. Biol.*, **243**, 327–344.
10. Thornton,J.M. and Gardner,S.P. (1989) Protein motifs and data-base searching. *Trends Biochem. Sci.*, **14**, 300–304.
11. Hendlich,M., Bergner,A., Gunther,J. and Klebe,G. (2003) Relibase: design and development of a database for comprehensive analysis of protein–ligand interactions. *J. Mol. Biol.*, **326**, 607–620.
12. 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.
13. Zhou,H.X. and Shan,Y. (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins*, **44**, 336–343.
14. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–741.
15. Pennec,X. and Ayache,N. (1998) A geometric algorithm to find small but highly similar 3D substructures in proteins. *Bioinformatics*, **14**, 516–522.
16. Kabsch,W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta Cryst.*, **A32**, 922–923.
17. Wang,G. and Dunbrack,R.L.,Jr (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.
18. Bonifati,V., Rizzu,P., van Baren,M.J., Schaap,O., Breedveld,G.J., Krieger,E., Dekker,M.C., Squitieri,F., Ibanez,P., Joosse,M. *et al.* (2003) Mutations in the DJ-1 gene associated with autosomal recessive early-onset Parkinsonism. *Science*, **299**, 256–259.
19. Huai,Q., Sun,Y., Wang,H., Chin,L.S., Li,L., Robinson,H. and Ke,H. (2003) Crystal structure of DJ-1/RS and implication on familial Parkinson's disease. *FEBS Lett.*, **549**, 171–175.
20. Hod,Y., Pentyala,S.N., Whyard,T.C. and El-Maghrabi,M.R. (1999) Identification and characterization of a novel protein that regulates RNA–protein interaction. *J. Cell Biochem.*, **72**, 435–444.
21. O'Brien,M.C. and McKay,D.B. (1995) How potassium affects the activity of the molecular chaperone Hsc70. I. Potassium is required for optimal ATPase activity. *J. Biol. Chem.*, **270**, 2247–2250.
22. Woehl,E.U. and Dunn,M.F. (1995) Monovalent metal ions play an essential role in catalysis and intersubunit communication in the tryptophan synthase bienzyme complex. *Biochemistry*, **34**, 9466–9476.
23. Suelter,C.H. (1970) Enzymes activated by monovalent cations. *Science*, **168**, 789–795.
24. Wells,C.M. and Di Cera,E. (1992) Thrombin is a Na^+-activated enzyme. *Biochemistry*, **31**, 11721–11730.
25. Dang,Q.D., Vindigni,A. and Di Cera,E. (1995) An allosteric switch controls the procoagulant and anticoagulant activities of thrombin. *Proc. Natl Acad. Sci., USA*, **92**, 5977–5981.
26. Di Cera,E., Guinto,E.R., Vindigni,A., Dang,Q.D., Ayala,Y.M., Wuyi,M. and Tulinsky,A. (1995) The Na^+x binding site of thrombin. *J. Biol. Chem.*, **270**, 22089–22092.
27. Di Cera,E. (2003) Thrombin interactions. *Chest*, **124**, 11S–17S.