doi:10.1093/nar/gki105

PDBSite: a database of the 3D structure of protein functional sites

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 August 18, 2004; Revised and Accepted October 15, 2004

ABSTRACT

The PDBSite database provides comprehensive structural and functional information on various protein sites (post-translational modification, catalytic active, organic and inorganic ligand binding, protein–protein, protein–DNA and protein–RNA interactions) in the Protein Data Bank (PDB). The PDBSite is available online at http://wwwmgs.bionet.nsc.ru/mgs/gnw/pdbsite/. It consists of functional sites extracted from PDB using the SITE records and of an additional set containing the protein interaction sites inferred from the contact residues in heterocomplexes. The PDBSite was set up by automated processing of the PDB. The PDBSite database can be queried through the functional description and the structural characteristics of the site and its environment. The PDBSite is integrated with the PDBSiteScan tool allowing structural comparisons of a protein against the functional sites. The PDBSite enables the recognition of functional sites in protein tertiary structures, providing annotation of function through structure. The PDBSite is updated after each new PDB release.

INTRODUCTION

Recognition of functional sites in proteins is a direct computational approach to the characterization of proteins in terms of biological and biochemical function. Over the years, it appeared that the three-dimensional (3D) coordinates can not only give hints on the protein's function, but also reveal it. Search for areas in the protein 3D structure showing structural similarity to the functional sites of other proteins, emerged full of promise. The repertoire of software tools kept growing to include the PROCAT database of the active site 3D templates (1), programs JESS (2) and TESS (3) for comparisons of the templates against the target protein; the PINTS program, which combines the database of functionally relevant patterns with a tool for local structural patterns search in the protein 3D structures (4). The ASSAM (5) and RIGOR (6) programs, among others, retrieve 3D motif matches.

Other approaches compare the target protein surface against a database of binding site surfaces. With Schmitt's approach (7), the putative binding sites are first automatically extracted from a structure, then characterized by descriptors in terms of physicochemical properties. Subsequently, the RELIBASE was developed to search for equivalents with regard to both shape and physicochemical properties (8). There exists a method for comparing the electrostatic surfaces of proteins against the eF-site database, with the expectation of identifying distantly related binding sites (9) and a more recent method for predicting nucleic acid binding sites relying on surface patches (10).

We have developed the PDBSiteScan program for fast comparisons of the protein 3D structure against the functional site database (11). The algorithm is efficient enough to enable high-performance search in real-time in databases containing known protein functional sites. The PDBSiteScan program compares the 3D protein structure using the backbone atoms (N, Cα and C). We chose this deliberately to avoid the possible structural differences between a site template and a target-protein due to ligand-induced conformational changes in a site template. Conformational differences can truly arise because it is common practice to use protein–ligand complexes for building binding site templates. Ligand-induced conformational changes are relevant to the prediction of protein–ligand interactions (12). We have shown that the PDBSiteScan approach is highly sensitive and specific in site recognition; thus, at certain values of the maximal distance mismatch (MDM) cut-off, sensitivity was 90% and specificity was almost 100% in cross-validation analysis of active site recognition in the hydrolase superfamily (11).

Here, we present a web-accessible PDBSite database of annotated protein functional sites. One goal was structural comparison of target protein against known functional sites, the other was to provide a flexible search for proteins on the basis of description of their functional sites. The PDBSiteScan program was integrated into PDBSite to reach the first goal.

*To whom correspondence should be addressed. Tel: +7 3832332971; Fax: +7 3832331278; Email: salix@bionet.nsc.ru

The physicochemical, structural and functional characteristics of the sites were included in the PDBSite, thereby providing the database retrieval. Efforts were made to offer the structural biologists opportunities to set up functional site samples meeting various criteria.

The environment of the sites can be important for their function (13). In fact, relationships between protein activities and physicochemical characteristics of functional site environment have been observed (14). It made sense to include the spatial environment calculated for the functional sites in PDBSite in the hope that the information would be useful for site recognition and delving into their function.

We applied PDBSite for searching novel functional sites in the mutants of the DNA-binding domain of the human protein p53 resulting from tumor-associated mutations. It was found that the G245C mutation gives rise to the zinc binding site, which partly overlaps the normal Zn^{2+} binding site. It was suggested that the extra Zn^{2+} binding site can compete with the normal site for zinc binding, thereby affecting the specificity of p53 binding to DNA. Quite obviously, PDBSite can be useful in tackling a wide range of tasks related to structural genomics and more specifically to structure-based drug design.

METHODS

The Protein Data Bank (PDB) fields HEADER, TITLE, KEYWDS, REMARK 800, SITE and ATOM were automatically processed to generate the PDBSite database. A description of the PDBSite database fields is available in the Supplementary Material. A PDB entry could contain data on a number of sites; if so, a separate entry was generated in the PDBSite to accommodate every one of the sites. To generate the description of proteins for site retrieval, the HEADER, TITLE and KEYWDS fields of the PDB were used. The textual descriptions of the site functions were extracted from the REMARK 800 field. A program was developed for data mining of the functional type of sites in the textual descriptions. Functional sites of ~180 types could be classified by the current version of the program. All sites that eluded our classification were referred to the Unclassified group.

Besides the sites retrieved from the PDB using the SITE field, the protein interaction sites with protein, RNA and DNA, inferred from the calculated contact residues in heterocomplexes, are regularly updated. The PDB heterocomplexes denoted by the word COMPLEX in the HEADER field were processed. We used contacts only between the different proteins for identifying the protein–protein interaction sites. The subunit contacts of the same protein were ignored in the current version of the PDBSite database. Biological and asymmetric units were not generated. The atomic coordinates defined in PDB were used without transformation. A residue having at least three atoms whose distance from any atom of the partner chain <5 Å was treated as a contact (15). The sites for protein– DNA and protein–RNA interactions were defined in a similar way.

The spatial environment of the sites was calculated by defining the residues in contact with the site residues. In this case, two residues were considered as contact if the distance between at least one atom pair, assigned to different residues, did not exceed 5 Å (16).

The spatial moment of the physicochemical property was calculated as

$$\mathrm{SM} = \left\{ \left[\sum_{i=1}^{N} h_i x_i\right]^2 + \left[\sum_{i=1}^{N} h_i y_i\right]^2 + \left[\sum_{i=1}^{N} h_i z_i\right]^2 \right\}^{1/2}, \qquad \mathbf{1}$$

where h_i ($i = 1, 2, \ldots, N$) is the value for the physicochemical property of the i-th site residue (e.g. hydrophobicity); N is number of site residues; x_i, y_i and z_i are the coordinates of the Cα atom of the i-th residue. The atomic coordinates was transformed relative to the geometric center of the site before calculating SM.

Site discontinuity in the protein primary structure was written as

$$\text{Discontinuity} = \frac{1}{N}\sum_{i=1}^{N}(p_{i+1} - p_i - 1), \qquad \mathbf{2}$$

where N denotes the site residue number, p_i is the order number of the i-th residue of the site in the protein sequence. The value expresses the average position number in primary structure between two neighbor site residues. To illustrate, if the site consists of a discontinuous sequence fragment, its estimated discontinuity index is 0.

DATABASE ACCESS

The database is available online at http://wwwmgs.bionet.nsc.ru/mgs/gnw/pdbsite/. The web page has a link to the Sequence Retrieval System (SRS) search in the PDBSite and a link to the PDBSiteScan program. SRS provides search tools that enable the user to retrieve functional sites using site characteristics. The PDBSiteScan searches by structural comparisons of a protein against a database functional site (11).

DISCUSSION

PDBSite statistics

The number of entries in the current release of the PDBSite database now stands at 10 204. The functional sites are subdivided into eight groups by their site description (see Table 1). Each group, in turn, is divided into subgroups by

Table 1. Grouping of functional site types in the PDBSite database

Groups of functional sites	Number of specific functions per group	Number of sites per group
Active[a]	222	1467
Post-translational modification	8	55
Ion metal binding	17	1506
Inorganic non-metal binding	9	657
Organic ligand binding	133	1130
Protein–protein interaction	995	1002
Protein–DNA interaction	1324	1329
Protein–RNA interaction	752	755
Pharmaceutical drug binding	14	28
Miscellaneous	—	2303

[a]The number of specific functions was estimated as the counts of unique EC enzymes from which the sites were extracted.

the specificity of site function. In addition to the eight subgroups, we singled out another group for the binding sites of pharmaceutical drugs. This group contains protein sites binding to organic compounds and proteins known to be medically relevant. The specific functions of the sites are given in the SITE_TYPE field (see the Supplementary Material). As shown in the table, the PDBSite contains a representative number of sites and it can be used for protein functional annotation. The PDBSite database actually contains a much greater number of functional site types than listed in Table 1. A substantial number of sites have eluded our classification and they were assigned to the Miscellaneous group in the table. Descriptions of site functions in the PDB database are often incomplete or poorly formalized. To do this better, additional information on their structure or their binding ligands is required. Nevertheless, these sites can be useful in protein annotation. Many of the sites contain descriptions that are not computer-readable, yet understandable to the user.

Annotation of mutant proteins

The annotation of mutant proteins with altered function resulting from natural mutations, e.g. single nucleotide polymorphism (17), is highlighted in structural genomics. It is axiomatic that genetic variation (mutation) is the basis behind both susceptibility to disease and response to drugs. The conventional approaches to function identification relying on search for homologous proteins are inapplicable to mutant protein annotation. With this in mind, we performed an analysis of tumorigenic mutations in the human p53 DNA-binding domain (DBD).

Mechanisms as to how tumorigenic DBD mutations undeniably cause reduction in p53 site-specific DNA-binding activity (18,19) were thought provoking: just by eliminating critical protein–DNA contact like R273H (20–22), by lowering thermodynamic stability to the level at which DBD is predicted to be for the most part unfolded, like R175H (23), or most likely by enhancing loss of the single bound Zn^{2+} ion (24,25).

In this analysis, we proceeded on the basic assumption that certain mutations can give rise to novel functional sites. The data on the p53 gene mutations were retrieved from the Swiss-Prot database [AC P04637, (26)]. The tertiary structures of the p53 mutants were calculated using the SWISS-MODEL server (27).

We found new extra functional sites in a number of mutant proteins. To illustrate, we revealed an extra zinc binding site that overlaps the normal zinc binding site in the mutant protein, G245C. According to the X-ray crystal structure of DBD (PDB ID 1gzh), Zn^{2+} is coordinated to C176, H179, C238 and C242 residues. The mutation G245C gives rise to a new site

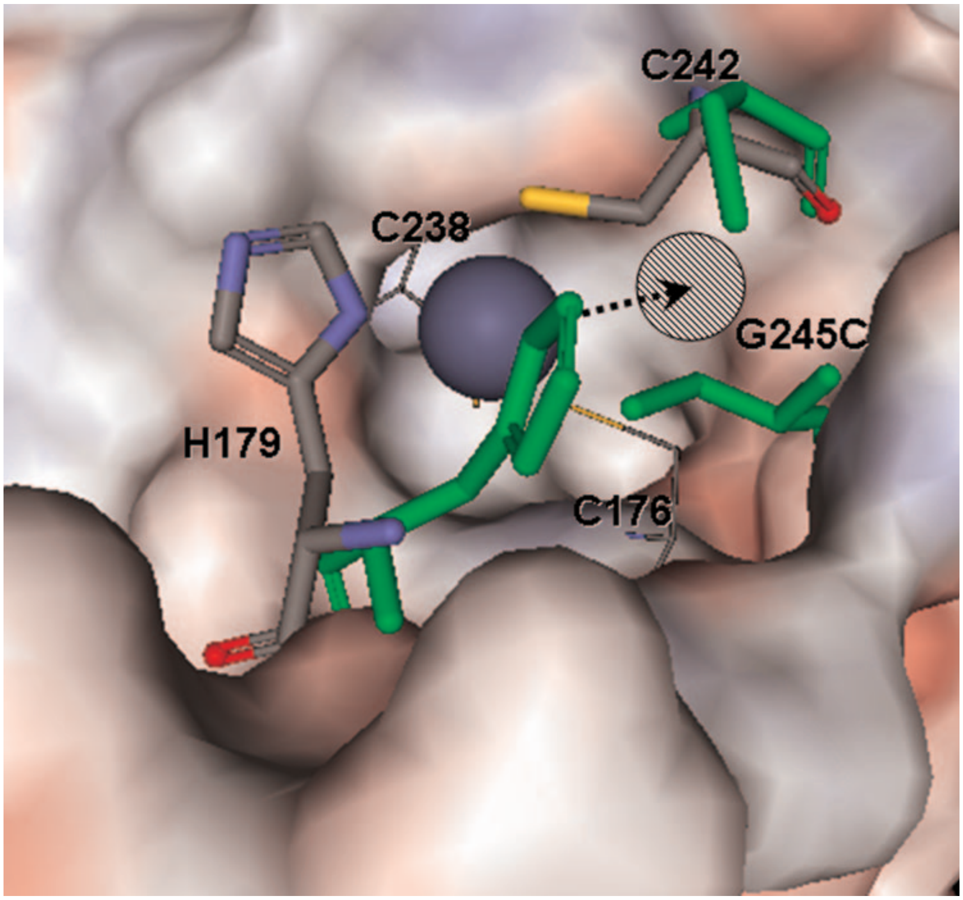

Figure 1. Superposition of the DNA-binding domain of human p53 (PDB ID 1gzh) to the Zn^{2+} binding site cytidine deaminase (PDB ID 1AF2). The cytidine deaminase site is highlighted in green. The normal position of the Zn^{2+} ion is shown as a blue-colored ball. The new potential position of the Zn^{2+}, in case an extra zinc binding site results from G245C mutation, is represented as a circle. The numbering of residues with their single-letter code is given for p53.

(H179, C242 and C245) similar in structure to the site for Zn^{2+} cytidine deaminase binding [PDB ID 1af2 (Figure 1)].

The mutation G245C was found in families with the Li-Fraumeni syndrome (28). On analysis of the functional significance of this germline mutation, it was concluded that malignant cells lose tumor-suppressor activity (29).

Our results suggest a molecular mechanism for the effect of the G245C substitution based on competition between normal and extra sites for Zn^{2+} binding. The molecular mechanism will remain conjectural until experimentally verified. If we are correct in regarding the novel site as extra Zn^{2+} binding, it can be a useful target in structure-based drug design.

PERSPECTIVES

The strategic consideration is: how to improve the method of functional site recognition? This will be a collaborative endeavour: our approach in conjunction with other approaches to functional site prediction. Benefits would be expected from broader approaches combining those based on structure alone, such as the exemplary THEMATICS (30,31), with those based on structural comparisons. We are striving to develop easy-to-use tools for the functional annotation of proteins when information on their homologs in scant. We envisage achieving this through incorporation; emphasis will be on integration of PDBSite into the protocols of functional protein annotation.

FUTURE DEVELOPMENTS

We intend to classify more completely the sites in the PDBSite database by their function. This will be achieved by expanding PDBSite to include information about ligands obtained through analyses of their complexes with proteins in the PDB; an approach based on structural comparisons of target sites against sites of clearly defined function will be implemented. A pharmaceutical description of drug and drug-like ligands that may have diagnostic or functional utility adapted from the data in the literature will be provided.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to A. N. Fadeeva for translating the manuscript from Russian into English. The work was partly supported by the Russian Fund for Basic Research (03-04-48506-a, 03-07-90181-b and 03-07-96833-p2003); the Siberian Branch of the Russian Academy of Sciences (Integration Project No. 119); Russian Ministry of Industry, Science and Technologies (43.073.1.1.1501); the US Civilian Research & Development Fund for the Independent States of the former Soviet Union (CRDF); Rup2-2629-NO-04; the Basic Research and Higher Education (BRHE) program NO-008-X1; Siberian Branch of the Russian Academy of Sciences (Project No. 10.4); and the State Contract of Federal agency on a science and innovations (01.106.11.0002).

REFERENCES

1. Wallace,A.C., Laskowski,R.A. and Thornton,J.M. (1996) Derivation of 3D coordinate templates for searching structural databases: application to the Ser-His-Asp catalytic triads of the serine proteinases and lipases. *Protein Sci.*, **5**, 1001–1013.
2. Barker,J.A. and Thornton,J.M. (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics*, **13**, 1644–1649.
3. 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.
4. Stark,A. and Russell,R.B. (2003) Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. *Nucleic Acids Res.*, **31**, 3341–3344.
5. Spriggs,R.V., Artymiuk,P.J. and Willett,P. (2003) Searching for patterns of amino acids in 3D protein structures. *J. Chem. Inform. Comp. Sci.*, **43**, 412–421.
6. Madsen,D. and Kleywegt,G.J. (2002) Interactive motif and fold recognition in protein structures. *J. Appl. Cryst.*, **35**, 137–139.
7. Schmitt,S., Kuhn,D. and Klebe,G. (2002) A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.*, **323**, 387–406.
8. 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.
9. Kinoshita,K., Furui,J. and Nakamura,H. (2001) Identification of protein functions from a molecular surface database, eF-site. *J. Struct. Funct. Genomics.*, **2**, 9–22.
10. Stawiski,E.W., Gregoret,L.M. and Mandel-Gutfreund,Y. (2003) Annotating nucleic acid-binding function based on protein structure. *J. Mol. Biol.*, **326**, 1065–1079.
11. Ivanisenko,V.A., Pintus,S.S., Grigorovich,D.A. and Kolchanov,N.A. (2004) PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. *Nucleic Acids Res.*, **32**, W549–W554.
12. Shoichet,B.K. and Kuntz,I.D. (1993) Matching chemistry and shape in molecular docking. *Protein Eng.*, **6**, 723–732.
13. Wei,L., Huang,E.S. and Altman,R.B. (1999) Are predicted structures good enough to preserve functional sites? *Structure*, **7**, 643–650.
14. Ivanisenko,V.A. and Eroshkin,A.M. (1997) Search for sites containing functionally important substitutions in series of related or mutant proteins. *Mol. Biol. (Moskow)*, **31**, 880–887.
15. Zhou,H.X. and Shan,Y. (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins*, **44**, 336–343.
16. Abagyan,R.A. and Totrov,M.M. (1997) Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. *J. Mol. Biol.*, **268**, 678–685.
17. Yoshida,A., Huang,I.Y. and Ikawa,M. (1984) Molecular abnormality of an inactive aldehyde dehydrogenase variant commonly found in Orientals. *Proc. Natl Acad. Sci. USA*, **81**, 258–261.
18. Kern,S.E., Pietenpol,J.A., Thiagalingam,S., Seymour,A., Kinzler,K.W. and Vogelstein,B. (1992) Oncogenic forms of p53 inhibit p53-regulated gene expression. *Science*, **256**, 827–830.
19. Unger,T., Nau,M.M., Segal,S. and Minna,J.D. (1992) p53: a transdominant regulator of transcription whose function is ablated by mutations occurring in human cancer. *EMBO J.*, **11**, 1383–1390.
20. Cho,Y., Gorina,S., Jeffrey,P.D. and Pavletich,N.P. (1994) Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science*, **265**, 346–355.
21. Bullock,A.N., Henckel,J., DeDecker,B.S., Johnson,C.M., Nikolova,P.V., Proctor,M.R., Lane,D.P. and Fersht,A.R. (1997) Thermodynamic stability of wild-type and mutant p53 core domain. *Proc. Natl Acad. Sci. USA*, **94**, 14338–14342.
22. Wong,K.B., DeDecker,B.S., Freund,S.M., Proctor,M.R., Bycroft,M. and Fersht,A.R. (1999) Hot-spot mutants of p53 core domain evince characteristic local structural changes. *Proc. Natl Acad. Sci. USA*, **96**, 8438–8442.

23. Bullock,A.N., Henckel,J. and Fersht,A.R. (2000) Quantitative analysis of residual folding and DNA binding in mutant p53 core domain: definition of mutant states for rescue in cancer therapy. *Oncogene*, **19**, 1245–1256.
24. Meplan,C., Richard,M.-J. and Hainaut,P. (2000) Redox signaling and transition metals in the control of the p53 pathway. *Biochem. Pharmacol.*, **59**, 25–33.
25. Butler,J.S. and Loh,S.N. (2003) Structure, function, and aggregation of the zinc-free form of the p53 DNA binding domain. *Biochemistry*, **42**, 2396–2403.
26. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
27. Schwede,T., Kopp,J., Guex,N. and Peitsch,M.C. (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.*, **31**, 3381–3385.
28. Malkin,D., Li,F.P., Strong,L.C., Fraumeni,J.F., Nelson,C.E., Kim,D.H., Kassel,J., Gryka,M.A., Bischoff,F.Z., Tainsky,M.A. and Friend,S.H. (1990) Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science*, **250**, 1233–1238.
29. Frebourg,T., Kassel,J., Lam,K.T., Gryka,M.A., Barbier,N., Andersen,T.I., Borresen,A.-L. and Friend,S.H. (1992) Germ-line mutations of the p53 tumor suppressor gene in patients with high risk for cancer inactivate the p53 protein. *Proc. Natl Acad. Sci. USA*, **89**, 6413–6417.
30. Ondrechen,M.J., Clifton,J.G. and Ringe,D. (2001) THEMATICS: a simple computational predictor of enzyme function from structure. *Proc. Natl Acad. Sci. USA*, **98**, 12473–12478.
31. Ringe,D., Wei,Y., Boino,K.R. and Ondrechen,M.J. (2004) Protein structure to function: insights from computation. *Cell. Mol. Life Sci.*, **61**, 387–392.