PDBSITE: A DATABASE ON PROTEIN ACTIVE SITES AND THEIR ENVIRONMENT

Ivanisenko V.A., Grigorovich D.A., Kolchanov N.A.

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
State Research Center of Virology and Biotechnology “Vector”, Koltsovo, Novosibirsk region, Russia

*Corresponding author

E-mail: salix@bionet.nsc.ru

Key words: biologically active protein sites, protein tertiary structure, databases

Resume

Motivation: The database Protein Data Bank (PDB) contains data on biologically active sites of many proteins: ligand-binding domains, enzyme catalytic centers, sites experiencing biochemical modification, etc. However, these data are of very limited access in the modern retrieval systems. Development of a database containing information on features of active sites and their spatial environment would provide a basis for comprehensive study of the properties of such sites.

Results: We have constructed the database PDBSite on biologically active sites retrieved from the database PDB. PDBSite contains description of site functions; lists of residues and their positions; structural features, calculated from 3D structures of the proteins; and physicochemical features of the sites and their spatial environment. The relationships between the properties of the residues of the sites and the residues of their environment have been analyzed.


Introduction

Information on biologically active sites is of paramount importance in solving many problems of molecular biology, biotechnology, and medicine. Highly specific biological activities of proteins are provided by the unique structure of active sites and their environment. For example, catalytic centers of enzymes occur in cavities (Chothia, 1976), and antigenic determinants are shaped as protrusions on a protein surface (Davies, Cohen, 1996). The structure of sites is largely dependent on their environment. The activity of many natural and mutant proteins depends on the physicochemical properties of the residues surrounding the functional sites (Ivanisenko, Eroshkin, 1997). Data on the 3D structures of proteins are necessary for determination of the spatial environment of their biologically active sites. The database PDB (Bernstein et al., 1997) contains the data on 3D structures of proteins. For many proteins, the amino acid residues composing biologically active sites are marked, and the sites are briefly described. The purpose of the present study is the development of a daughter database PDBSite on the specific features of spatial organization of the biologically active sites stored in PDB and their environment.

For this purpose, we developed original methods, algorithms and software for calculating the 3D environment of sites and their structural and physicochemical properties.

Most investigations of the structure–function organization of functional sites are aimed at the invariant properties of these sites or their environment (Bagley, Altman, 1995; Sekharudua, Sundaralingam, 1988). We performed search for correlations among the variable properties of sites and their environment. This showed that a correlated change of the property pairs “site–environment” is a typical feature of the sites. These results may contribute to understanding the mechanisms of site operation and evolution.

Materials and Methods

The PDBSite database includes data obtained by processing of the following PDB fields: HEADER, TITLE, KEYWDS, REMARK 800, SITE, and ATOM. Grammar analysis programs were used for processing of PDB records. If a single PDB record contained data on several sites, individual records were created for each site in PDBSite. The Internet access to PDBSite was performed with the use of the Sequence Retrieval System (SRS).

The spatial environment of sites was calculated as follows. The least parallelogram including all the atoms of the amino acid residues of a site was constructed from the coordinates of the protein atoms. The spatial environment of the site was assumed to include each amino acid residue such that at least one of its atoms was within the parallelogram.

The following properties of sites and environment were calculated: solvent accessibility of amino acid residues; the mean value, sum, and spatial moment of amino acid physicochemical indices; coordinates of the geometrical center of mass for each residue; and the pairwise distances between the centers of mass of the residues. Another site index calculated was its discontinuity in the primary structure. The spatial moment was calculated as follows:
\[ SM = \sqrt{\left( \sum_{i=1}^{N} p_i x_i \right)^2 + \left( \sum_{i=1}^{N} p_i y_i \right)^2 + \left( \sum_{i=1}^{N} p_i z_i \right)^2} \]

where \( p_i (i = 1, 2, ..., N) \) is the value of a certain property of the \( i \)th residue of the 3D site, comprising \( N \) amino acids; \( x_i, y_i, z_i \) are coordinates of the Ca atom of the \( i \)th residue taken with reference to the geometrical center of the 3D site.

The discontinuity index of a site was taken to be
\[ \frac{1}{N} \sum_{i=1}^{N} (P_{i+1} - P_i - 1) \]
where \( N \) is the number of residues of the site and \( P_i \) is the ordinal number of the \( i \)th residue of the site in the protein sequence. This index reflects the mean number of positions between the neighboring residues of the site in the primary structure. For example, if a site consists of a continuous sequence fragment, then the discontinuity index is zero. To calculate the solvent accessibility of amino acids, we used the program DSSP (Kabsch, Sander, 1983).

### Results and Discussion

The fields of the PDBSite database used for search in the system SRS are listed and briefly described in Table 1.

<table>
<thead>
<tr>
<th>Field name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Entry identifier.</td>
</tr>
<tr>
<td>PDBID</td>
<td>PDB ID code.</td>
</tr>
<tr>
<td>Header</td>
<td>PDB classification for the entry. The field content corresponds to that in PDB.</td>
</tr>
<tr>
<td>Title</td>
<td>Title for experiment or analysis described in the entry. The field content corresponds to that in PDB.</td>
</tr>
<tr>
<td>Keyword</td>
<td>Keywords describing the macromolecule. The content corresponds to that of KEYWDS in PDB.</td>
</tr>
<tr>
<td>Molecule</td>
<td>Contains names of macromolecules from the COMPND of PDB and is designed to look for entries by the names of macromolecules.</td>
</tr>
<tr>
<td>NumSiteChains</td>
<td>Number of different chains to which the residues of the site belong.</td>
</tr>
<tr>
<td>SiteDescr</td>
<td>Description of the site. The content corresponds to that of SITE_DESCRIPTION subfield of REMARK 800 field of PDB.</td>
</tr>
<tr>
<td>ResidueNotAA</td>
<td>Names of residues that are not amino acids but are present in the site.</td>
</tr>
<tr>
<td>LenSite</td>
<td>Number of residues in the site.</td>
</tr>
<tr>
<td>LenSurround</td>
<td>Number of residues in the site environment.</td>
</tr>
<tr>
<td>ExposureSite</td>
<td>Average exposure of residues in the site.</td>
</tr>
<tr>
<td>ExposureSurround</td>
<td>Average exposure of residues in the site environment.</td>
</tr>
<tr>
<td>Discontinuity</td>
<td>Discontinuity of the site according to its primary structure.</td>
</tr>
</tbody>
</table>

In addition, the database contains data on the structural and physicochemical indices of sites and their spatial environment that are not used for search but can be applied to analysis of the structure-functional organization of sites.

The PDBSite base contains the descriptions of 4723 sites. For statistical analysis, we obtained a nonredundant set by exclusion of complete analogs related to protein duplication in PDB. This set included 4038 sites. All the sites can be divided into three groups: continuous, discontinuous, and such discontinuous sites whose residues correspond to different subunits of a molecule. The proportion of the last group was considerable - 10.6%. The distribution of sites for the discontinuity index is shown in Fig. 1.

The figure shows that most sites present in the database are discontinuous. Thus, the patterns based only on the primary structure can hardly be developed for many sites. Recognition of the residues of such sites requires at least consideration of the matrix of distances between protein residues in the tertiary structure. We found that sites with increased discontinuity have limited solvent accessibility (Fig. 2); i.e., burred residues form highly discontinuous sites.

For further analysis, sites were grouped according to their functions. Only those sites whose function is unambiguously described in the field SiteDescr were taken into consideration. Of them, we chose site types including no less than 10 members. Thus, 3611 sites of 30 types were examined. Figure 3 shows the hierarchical classification of sites of different types for their mean amino acid composition. The sites clustered into two major groups in the resulting tree. One of them is dominated by organic ligand-binding sites, and the other, by metal ion-binding sites. Sites of acetylation fell into the first group, and the sites of glycosylation and phosphorylation, to the second one. Catalytic centers of enzymes belong to the first group. Both groups contain also local clusters of sites of different types. In general, the tree constructed according to the amino acid composition of sites is in good agreement with their function.
Features presented in the PDBSite database were used for analysis of correlations between the physicochemical properties of sites and their environment.

Figure 4 illustrates the correlation between the hydrophilicity of the Mn-binding, Co-binding, and Mg-binding sites, pooled into one group, and the hydrophobicity of their solvent-exposed environment, i.e., residues on the molecule surface. It is apparent that the hydrophobicity of the exposed environment increases with site hydrophilicity.

The results of our analysis suggest that the function of site environment manifests itself as early as the stage of initial recognition of the site target and correct orientation of the site with respect to the target.

The mean or total physicochemical properties of sites were found to correlate with their spatial moment. This can be related to the importance of the distribution of these properties over the spatial structure of the site.

In particular, the increase in the number of positively charged residues in ATP-binding sites correlates with the increase in charge moment (Fig. 5). To put it differently, residues are aggregated into clusters of positively and negatively charged ones.
Conclusions

We have developed the PDBSite database, which can be applied to various tasks in the investigation of the functions of proteins and their active centers.

The PDBSite database is equipped with an automated system for search for structural similarity between active sites and a user-defined 3D structure of a certain protein (Ivanisenko et al., 2002). The search for structural similarity over PDBSite allows recognition of active sites in the spatial structure of proteins.

Our results indicate that the correlations between physicochemical properties of sites and their spatial environment are characteristic of globular proteins. A more complete solution of the problem of finding relations between site properties and environment demands further analysis, involving the conformation and physicochemical properties of sites and their environment. Application of methods of molecular dynamics and conformation analysis is promising for further study of this problem.

Acknowledgements

The study was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376 and 01-07-90084); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Project № 65); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049).

References