



A tool to assist the study of specific features at protein binding sites

Víctor Santander, Miguel Angel Portales and Francisco Melo*

Pontificia Universidad Católica de Chile, Departamento Genética Molecular y Microbiología, Facultad de Ciencias Biológicas, Alameda 340, Santiago, Chile

Received on January 6, 2003; accepted on February 20, 2003

ABSTRACT

The Protein Data Bank contains a large amount of proteins that have been solved with small ligands bound to them. This constitutes a rich source of information for the study of the specific requirements of protein sites to bind small molecules with a favorable free energy. The specific atomic composition and three-dimensional geometric restraints of protein binding sites for different ligands could be easily obtained from there. The development of accurate binding site descriptors in proteins constitutes a valuable tool to assist in the large-scale prediction and annotation of protein function in whole genomes. In this work, an integrated database containing some processed and calculated protein/ligand information is described. It is expected that this database will constitute a useful tool for people working in the prediction of protein function from its structure. The database is accessible from the Internet through a web server located at: <http://protein.bio.puc.cl>

Contact: fmelo@genes.bio.puc.cl

Keywords: Ligands, function, structure, proteins, binding sites, conformational analysis

INTRODUCTION

Little is known about the atomic configuration and conformational requirements of native proteins in order to bind small molecule ligands with a favorable free energy under standard conditions. A large amount of experimental data is already available where thousands of proteins have been co-crystallized with small ligands bound to them. This data is growing rapidly and continuously as more proteins are being deposited daily in the Protein Data Bank (PDB, Berman *et al.*, 2002). Thus, it is necessary to develop new tools to extract those protein binding sites in order to begin their chemical and geometrical characterization. This constitutes the first step in the development of protein binding site predictors that can help in the prediction and annotation of protein function. This is particularly relevant to those methods that are expected to be applied in a large scale, which due to the large number of proteins

available cannot be based in complex molecular docking simulations.

In this work, we present a novel database that contains the whole PDB integrated with the CATH database (Orengo *et al.*, 1997; Pearl *et al.*, 2000) and with several calculated subsets of protein atoms that are in direct contact with the ligands, as well as the ligand conformations themselves. This database is intended to answer the following questions:

1. How many occurrences of a given ligand are in the PDB?
2. Which proteins is a given ligand binding to?
3. How many different folds bind a given ligand?
4. Which specific folds is a given ligand binding to?
5. In what proportion a ligand binds one or more folds?
6. Which protein atoms are binding to a ligand at different distance cutoffs?
7. What is the spatial distribution of the protein atoms at the ligand binding site?
8. How conserved or variable are the observed conformations of a ligand in its bound state to proteins?

The database is freely accessible from the Internet through a web server located at: <http://protein.bio.puc.cl>

METHODS

The database is implemented in MySQL with a PHP engine and a set of HTML forms for different purposes. It is expected that the total number of available forms will increase with time based on the future feedback from the users. The database contains a total of eleven linked tables; six of them containing processed CATH data, one containing crude PDB general information data, and four containing PDB processed data. A detailed description of the tables, along with their field types and contents, is available on-line. The web server can provide partial and complete PDB files, thus if the web client is properly set, three-dimensional coordinates of

*To whom correspondence should be addressed.

ligands and proteins can be displayed graphically through an external software such as RasMol (Sayle and Milner-White, 1995) or an installed plug-in like CHIME by MDL Information Systems, Inc (<http://www.mdlchime.com>). It is also possible to generate *.tar.gz* files containing the three-dimensional atomic coordinates of all the ligands found, along with the corresponding coordinates of the atoms that constitute the protein binding site for each ligand.

Four programs were written in ANSI C language to build this database. The first one used as input the raw domain file release 2.4 from CATH database (Orengo *et al.*, 1997; Pearl *et al.*, 2000) generating the data for the six tables in this database that contain CATH information, which were then imported into MySQL. The second program used as input the whole list of proteins from the PDB, identifying the different ligands, storing the three-dimensional coordinates of them separately into disk in PDB format, calculating the centroids and the maximal internal pairwise atomic Euclidean distance for each ligand, and assigning a unique identifier to each ligand based on a sequential numbering. The data generated by this program was then imported into MySQL and also used to feed the other programs. The third program used as input the whole PDB list of proteins and the ligand information from the output of the previous program, splitting the information in a chain-by-chain basis, and selecting the contacts between ligands and protein chains at different distance cutoffs of 3.0, 3.5, 4.0, 4.5, and 5.0 Angstroms. The protein atoms in contact with the ligands at the different distance cutoffs were stored to the disk in PDB format using the unique ligand identifier, thus they can be easily obtained once needed. Finally, the fourth program used CATH, PDB, and the output from the other programs to calculate the total number of contacts that a given ligand had with a particular fold at the different distance cutoffs. Thus, a fractional index of fold contacts for any ligand could be calculated and incorporated into the database.

This database will be updated on a regular basis every three months with each new CD-ROM release from the PDB.

RESULTS

The initial database was derived from a total of 16,977 proteins (26,272 protein chains) that were available in the PDB (July 2002 release). Out of these, 10,666 (63%) proteins (18,065 protein chains) contained at least one small ligand bound to them. A total number of 42,363 ligands were extracted from there. Sixty eight percent of the ligands were bound to one or more chains that had a CATH annotation. It was possible to identify 2,972 different ligands, where only about five percent of them had more than 50 observations in the database. It is expected that this database will grow continuously in time with each new release of the PDB, allowing the characterization of the specific features for a growing number of new protein binding sites.

ACKNOWLEDGEMENTS

This work was funded by projects from Fundación Andes (# 13600/4) and FONDECYT (# 1010959).

REFERENCES

- Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr. D*, **58**, 899–907.
- Bernstein, H.J. (2000) Recent changes to RasMol, recombining the variants. *Trends in Biochemical Sciences*, **25**, 453–455.
- Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
- 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.
- Sayle, R. and Milner-White, E.J. (1995) RasMol: Biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.