*D306–D309* *Nucleic Acids Research, 2006, Vol. 34, Database issue*
*doi:10.1093/nar/gkj105*

# The PMDB Protein Model Database

**Tiziana Castrignanò, Paolo D'Onorio De Meo, Domenico Cozzetto[1], Ivano Giuseppe Talamo[1] and Anna Tramontano[1,2,*]**

CASPUR, Consorzio Interuniversitario per le Applicazioni di Supercalcolo per Università e Ricerca, Via dei Tizii, 6/b, I-00185 Rome, Italy, [1]Department of Biochemical Sciences, University 'La Sapienza', P.le Aldo Moro, 5, I-00185 Rome, Italy and [2]Istituto Pasteur—Fondazione Cenci Bolognetti, University 'La Sapienza', P.le Aldo Moro, 5, I-00185 Rome, Italy

Received July 28, 2005; Revised and Accepted October 17, 2005

## ABSTRACT

**The Protein Model Database (PMDB) is a public resource aimed at storing manually built 3D models of proteins. The database is designed to provide access to models published in the scientific literature, together with validating experimental data. It is a relational database and it currently contains >74 000 models for ~240 proteins. The system is accessible at http://www.caspur.it/PMDB and allows predictors to submit models along with related supporting evidence and users to download them through a simple and intuitive interface. Users can navigate in the database and retrieve models referring to the same target protein or to different regions of the same protein. Each model is assigned a unique identifier that allows interested users to directly access the data.**

## INTRODUCTION

The data deluge brought about by the genomic projects has fostered an unprecedented level of expectation for new medical, pharmacological, environmental and biotechnological discoveries. Proteins mediate the majority of the functions of an organism, and all these functions are, by and large, determined by the proteins' 3D structure.

Despite the progress achieved so far by structural genomics projects (1), the exploration of the complete protein structure space through experimental techniques such as X-ray crystallography and NMR spectroscopy is still out of reach, because these techniques are time and resource consuming and not necessarily successful in all cases. Consequently the gap between the numbers of known protein structures and sequences is steadily increasing.

Natural proteins spontaneously assume a unique 3D structure that, by and large, only depends upon the protein sequence. The problem of understanding the rules governing the folding process is very challenging and as yet unsolved. However, approximate methods for inferring the structure of a protein from its amino acid sequence are flourishing (2). Their results are of enormous relevance in many fields, from medicine to biology, from biotechnology to pharmacology. Information derived from protein models has indeed proven to be useful by itself and in combination with experiments. Protein models have been shown to be instrumental for the refinement of experimental structures (3,4), the design of site-directed mutants (5), the characterization of molecular function (6) and structure-based drug design (7).

Not surprisingly, a growing number of scientific papers reporting the results of modelling experiments and their application to the design and interpretation of experiments are appearing in the literature. Unfortunately, the models described in these reports are rarely publicly available and, in general, only accessible via direct interaction with the authors. The difficulty of mining the available structural model data leads to duplication of efforts and impairs the possibility of numerically evaluating the correctness of the models when the experimental structure becomes available.

The establishment of public repositories for these protein 3D models can partly overcome these problems. Specialized databases, such as ModBase (8) and the SWISS-MODEL repository (9), are already available for automatically built protein structure models. We have developed, and describe here, a Protein Model Database (PMDB) where manually built models can be deposited and retrieved, together with their supporting information.

## DATABASE CONTENT AND WEB ACCESS

PMDB (interactively accessible at http://www.caspur.it/PMDB) is a relational database of protein models submitted by users and obtained with different structure prediction techniques. The database is implemented on a Linux server (Suse Enterprise Server 9) running Apache, and the management system is MySQL 4.1.12. The queue management system is

*To whom correspondence should be addressed. Tel: +39 0649910556; Fax: +39 0649910717; Email: anna.tramontano@uniroma1.it

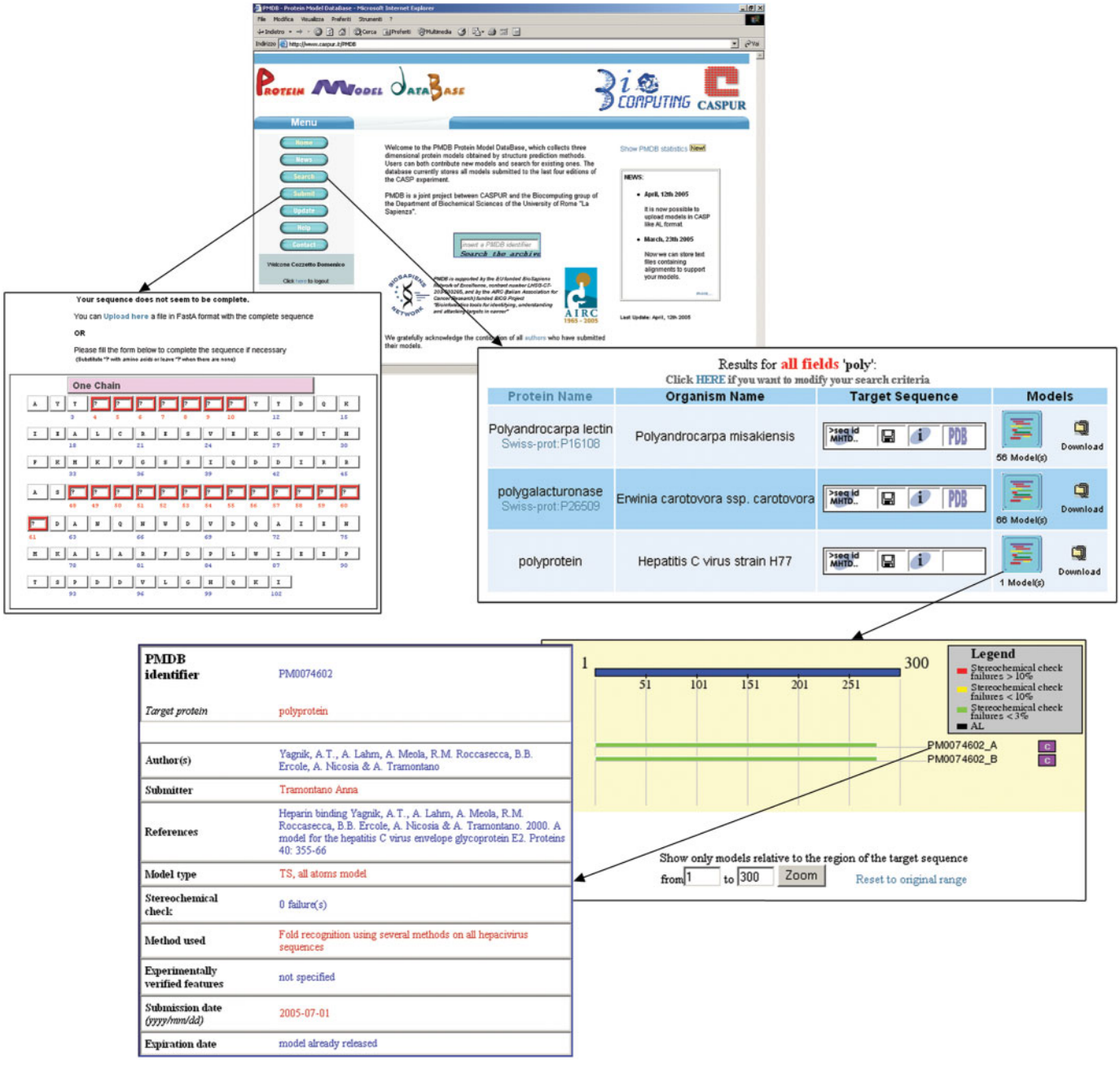


**Figure 1.** PMDB overview. Information about each of the models satisfying the search criteria can be easily retrieved. When a user uploads a model, its amino acid sequence is automatically retrieved from its coordinate file. Residues for which coordinates are not available in the PDB model file, if any, can be manually inserted.

written in Perl. PHP scripts and GD libraries are used for launching applications such as Blast and for display, respectively.

The current release contains >74 000 models for ~240 proteins, the majority of which are predictions submitted to the 'Critical Assessment of Techniques for Structure Prediction' experiment (2). Other models include those generated by our group (10,11) and models that we uploaded using published alignments (12–15).

The database entry point is a protein target, for which one or more structural models can be present in the database. Available information for each target includes the protein name, sequence and length, organism and, whenever applicable, links to the SwissProt sequence database (16). Several models can be present for each target protein, or for different regions of the same target protein and the user can navigate through them using a graphical view shown in Figure 1. After the structure of a target is solved, the database entry is also linked to the experimental structure in the PDB (17).

Models can be submitted in the form of a PDB file (TS format) or as an alignment to one or more known protein structures (AL format). In the latter case the coordinates of the backbone of the model are built using the AL2TS program (http://predictioncenter.org/local/al2ts/al2ts.html).

When a user submits a model for a protein, the system verifies whether the target already exists (i.e. there is already a model for some regions of the same protein). If not, the target is created and the model mapped to it. Unless the target is an artificial or mutant protein, the target entry is linked to existing sequence databases [at present the NCBI nr database (18)]. The predictor can provide the NCBI id of the protein (in which case the system performs a sequence check), ask for a BLAST search in the database to retrieve the id (if more than one entry matches the sequence, the user is requested to select the correct one), or inform the system that the target is not expected to be present in any sequence database.

The sequence of the target is derived from the submitted model PDB or alignment file. In the former case, if the distance between consecutive Cα is larger than expected for connected residues, the user is asked whether he or she wants to complete the target sequence (Figure 1). The system also reports cases where atoms in the model are closer than the sum of their van der Waals radii.

The database stores information about the author of the model, a short description of the method used and supporting evidence, in the form, for example, of a multiple sequence alignment. Submitters are also asked to assign a reliability value to their model(s) and a literature reference that can also be provided at a later stage. Models can be kept on hold upon request and made available to the general users after at most 6 months from deposition.

At the end of the submission procedure, the model is assigned a unique identifier.

The user interface allows the model(s) to be searched by protein or organism name, protein accession number (in the nr database), author, PMDB model identifier, model type (i.e. a complete coordinate set indicated by TS or an alignment to a known structure indicated by AL). It is also possible to perform sequence similarity searches via BLAST (19). Search results are displayed in the form of a table, listing the records satisfying all selected criteria (Figure 1). Each row refers to a target sequence and related models, along with summary information. Every model that is not on hold can be downloaded or displayed through the 3D visualization program RASMOL (20).

## FUTURE DEVELOPMENTS

Immediate future plans for the database include the possibility of using UNIPROT identifiers (21) for the protein targets and to perform more sophisticated searches.

We also plan to add provisions for evaluating the models, other than the simple stereochemical checks performed at present, using tools such as WHATCHECK (22), Verify3D (23) and PROSA (24) as well as tools to automatically evaluate the quality of models of proteins the structure of which is subsequently solved (25). This will permit, in the future, to analyse the correlation between the actual quality of the models with the reliability values assigned by the authors and with those estimated by automatic verification tools.

## ACKNOWLEDGEMENTS

The authors are grateful to Dr K. Fidelis for providing the AL2TS source code. PMDB is supported by the EU funded BioSapiens Network of Excellence, contract number LHSG-CT-203-503265, and by the AIRC funded BICG Project 'Bioinformatics tools for identifying, understanding and attacking targets in cancer'. Funding to pay the Open Access publication charges for this article was provided by the European Union.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Todd,A.E., Marsden,R.L., Thornton,J.M. and Orengo,C.A. (2005) Progress of structural genomics initiatives: an analysis of solved target structures. *J. Mol. Biol.*, **348**, 1235–1260.
2. Moult,J., Fidelis,K., Zemla,A. and Hubbard,T. (2003) Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins*, **53** (Suppl. 6), 334–339.
3. Giorgetti,A., Raimondo,D., Miele,A.E. and Tramontano,A. (2005) Evaluating the usefulness of protein structure models for molecular replacement. *Bioinformatics*, **21** (Suppl. 2), 72–76.
4. Jones,D.T. (2001) Evaluating the potential of using fold-recognition models for molecular replacement. *Acta Crystallogr. D*, **57**, 1428–1434.
5. Peitsch,M.C. (2002) About the use of protein models. *Bioinformatics*, **18**, 934–938.
6. Laskowski,R.A., Watson,J.D. and Thornton,J.M. (2003) From protein structure to biochemical function? *J. Struct. Funct. Genomics*, **4**, 167–177.
7. Jacobson,M. and Sali,A. (2004) Comparative protein Structure Modelling and its applications to drug discovery. *Annu. Rep. Med. Chem.*, **39**, 259–274.
8. Pieper,U., Eswar,N., Braberg,H., Madhusudhan,M.S., Davis,F.P., Stuart,A.C., Mirkovic,N., Rossi,A., Marti-Renom,M.A., Fiser,A. *et al.* (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.*, **32**, D217–D222.
9. Kopp,J. and Schwede,T. (2004) The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Res.*, **32**, D230–D234.
10. Procopio,M., Lahm,A., Tramontano,A., Bonati,L. and Pitea,D. (2002) A model for recognition of polychlorinated dibenzo-p-dioxins by the aryl hydrocarbon receptor. *Eur. J. Biochem.*, **269**, 13–18.
11. Yagnik,A.T. *et al.* (2000) A model for the hepatitis C virus envelope glycoprotein E2. *Proteins*, **40**, 355–66.
12. Townley,H.E., Sessions,R.B., Clarke,A.R., Dafforn,T.R. and Griffiths,W,T. (2001) Protochlorophyllide oxidoreductase: a homology model examined by site-directed mutagenesis. *Proteins*, **44**, 329–335.
13. Flanagan,J.U., Rossjohn,J., Parker,M.W., Board,P.G. and Chelvanayagam,G. (1998) A homology model for the human theta-class glutathione transferase T1-1. *Proteins*, **33**, 444–454.
14. Brinkworth,R.I., Fairlie,D.P., Leung,D. and Young,P.R. (1999) Homology model of the dengue 2 virus NS3 protease: putative interactions with both substrate and NS2B cofactor. *J. Gen. Virol.*, **80**, 1167–77.
15. Thomas,B.A., Church,W.B., Lane,T.R. and Hammock,B.D. (1999) Homology model of juvenile hormone esterase from the crop pest, *Heliothis virescens*. *Proteins*, **34**, 184–196.
16. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
17. 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.
18. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–D45.
19. Altschul,S., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

20. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci*., **20**, 374–376.
21. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*., **32**, D115–D119.
22. Hooft,R., Vriend,G., Sander,C. and Abola,E.E. Errors in protein structures. *Nature*, **381**, 272–272.
23. Eisenberg,D., Luthy,R. and Bowie,J.U. (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol*., **277**, 396–404.
24. Sippl,M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins*, **17**, 355–362.
25. Zemla,A., Venclovas,C., Moult,J. and Fidelis,K. (1999) Processing, and analysis of CASP3 protein structure predictions. *Proteins*, Suppl. 3, 22–29.