BIOINFORMATICS APPLICATIONS NOTE
Vol. 20 no. 15 2004, pages 2482–2484
doi:10.1093/bioinformatics/bth262

The PDB-Preview database: a repository of *in-silico* models of 'on-hold' PDB entries

*Daniel Fischer[1], Jakub Paś[2] and Leszek Rychlewski[3,]**

[1]Buffalo Center of Excellence in Bioinformatics, University of Buffalo, 901 Washington Street Ste. 300, Buffalo, NY 14203, USA, [2]Department of Physics, Adam Mickiewicz University, ul.Umultowska 85, 61-614 Poznań, Poland and [3]BioInfoBank Institute, ul. Limanowskiego 24A, 60-744 Poznań, Poland

Received on October 7, 2003; revised on February 3, 2004; accepted on March 26, 2004
Advance Access publication April 8, 2004

ABSTRACT

Summary: The PDB-Preview database is a dynamic web repository of *in-silico* predicted three-dimensional (3D) models of experimentally determined structures that are deposited into the PDB but are not yet publicly released, and are kept 'on-hold'. The PDB-Preview database is automatically generated on a weekly basis by the bioinfo.pl meta-server, which uses top-of-the-line fold-recognition methods. The PDB-Preview provides biologists with preliminary fold assignments well before the experimentally determined 3D structures are released.

Availability: http://bioinfo.pl/PDB-Preview/

Contact: leszek@bioinfo.pl

INTRODUCTION

Knowing a protein's three-dimensional (3D) structure is often essential for biological research. The Protein Data bank (PDB) (Sussman *et al*., 1998) is a repository of 3D structures of proteins and nucleic acids and is an essential resource for structural biologists. However, not all the entries in the PDB are publicly available. A new structure can be deposited as an 'on-hold' entry, which means that the 3D coordinates of the protein are not available to the public before final release.

Fortunately, in many cases, relatively accurate computational models can be generated for those entries with no available 3D structure, using homology modeling and fold-recognition (FR) methods (Kinch *et al*., 2003; Rychlewski *et al*., 2003). Until recently, there were no fully-automated, easy-to-use, reliable FR methods available for a non-expert user. Thus, many biologists would not invest their valuable time attempting to generate a computational model for a given protein, even if relatively accurate models could be obtained. The latest attempts to evaluate FR methods (Kinch *et al*., 2003; Rychlewski *et al*., 2003) show that recently developed fully-automated meta-servers (Bujnicki *et al*., 2001) greatly increase the prediction accuracy. Thus, the creation of a database of predicted FR models for those on-hold entries that cannot be obtained with standard homology modeling tools became possible, and biologists can now, in a matter of seconds, access this database to download the generated computational models.

DATABASE STRUCTURE

Every week, the newly deposited on-hold entries in the PDB are scanned to select those that correspond to proteins with no significant sequence similarity to any protein of known structure. Each such entry is processed by the bioinfo.pl meta-server. The meta-server compiles FR results from a large number of servers world-wide and computes a 'consensus' prediction using the 3D-Jury method (Ginalski *et al*., 2003). Each prediction is automatically incorporated into the PDB-Preview database available at http://bioinfo.pl/PDB-Preview. On July 2003, the database contained predictions for a total of just over a hundred on-hold entries.

The 3D-Jury method is a sensitive and reliable consensus method, which has proved to be among the best fully automated prediction methods in the last prediction assessment experiments CASP5 and CAFASP3 (Fischer *et al*., 2003). The individual methods used by 3D-Jury are listed at http://bioinfo.pl/Meta/servers.html. The PDB-Preview provides three basic prediction parameters that give the users an indication of the reliability of the collected models: (1) a Blast (Altschul *et al*., 1990) E-value of the best hit to a PDB entry, which has to be above 0.001 in order to be included in the database. Targets with E-values below this threshold are considered homology-modeling targets. This does not mean that all homology-modeling targets can be easily modeled, because in many cases some aligned regions can be wrong, or parts of the protein can be left unaligned. Nevertheless, this filtering procedure is employed for simplicity. (2) A PDB-Blast E-value of the best hit to a PDB protein (obtained with a sequence profile comprised after four iterations of PSI-Blast (Altschul *et al*., 1997) on the nr databases clustered at 70% sequence identity using the CD-HIT program (Li *et al*., 2001)

*To whom correspondence should be addressed.

and masked with low complexity filters). (3) The 3D-Jury score obtained with default parameters. This score reflects the consistency of the models collected from the structure prediction servers (Koh *et al*., 2003; Rychlewski *et al*., 2003). From extensive tests on targets with known structures, it has been found that 3D-Jury predictions with scores above 50 correspond to essentially correct predictions, meaning that the overall folds of the predicted models are structurally similar to the corresponding experimental structures in over 90% of the cases.

In the PDB-Preview page, the PDB-Blast E-value is displayed in yellow if it is below 0.002 (35 cases in July 2003). This indicates a simple FR target (by no means a trivial prediction). In most cases, this is confirmed by a high 3D-Jury score (above 50), which is also shown in yellow. If the PDB-Blast prediction is not confirmed by 3D-Jury then the 3D-Jury score is displayed in red (four cases). In most cases, this indicates technical problems such as PDB-Blast and 3D-Jury scores just opposite the cut-off values, or spurious significant PDB-Blast scores (e.g. a short histidine tag in *Thermotoga maritima* 1070 protein 1nc7). In addition, the PDB-Preview highlights with blue those 3D-Jury predictions that are regarded as confident but cannot be obtained with a significant score by PDB-Blast. These probably entail the most interesting prediction targets, and 26 such cases are reported in July 2003. The most confident non-trivial predictions obtained for already released protein structures are shown in Table 1. Manual inspection of the results indicates that the predicted folds are essentially correct.

Table 1. Results for models of PDB-Preview targets assessed by the 3D-Jury as confidently correct (3D-Jury score >50)

PDB code	3D-Jury score	Correct residues	
		3D-Jury	PDB-Blast
1nw1	139	103	89
1m3u	126	115	41
1mki	123	75	25
1ok4	122	78	50
1nc5	108	112	30
1nof	102	87	18
1pzd	94	59	14
1qwg	92	80	36
1omx	92	75	9
1q77	80	51	23
1o0i	80	60	11
1ow4	73	74	11
1m4o	73	53	44
1ixl	72	72	45
1ni9	64	77	9
1nf2	63	83	48
1ox7	61	63	29
1o7b	59	58	7
1q1h	58	42	47
1otk	57	109	17
1uaq	57	63	29
1p5h	56	43	11
1p5h	56	43	11
1oz9	53	51	19
1o9h	53	81	8
1ucp	52	63	15
1q48	50	45	16

In all models the numbers of correctly predicted residues (within 3A from the native conformation) is above 40. The last column shows the number of correctly predicted residues that models obtained with PDB-Blast would obtain. Most PDB-Blast models correspond to incorrect predictions (using templates of unrelated folds or incorrect alignments) and in all cases but one (1q1h), the 3D-Jury models have higher quality than the models obtained with PDB-Blast.

QUALITY CONTROL

Our system periodically removes from the PDB-Preview database those PDB entries whose status has changed from on-hold to available. At this point, the accuracy of the stored predictions can be evaluated, because the experimental 3D coordinates become available. Thus, a second section of the PDB-Preview database corresponds to predictions of previously on-hold entries that have been released. These are available at http://bioinfo.pl/PDB-Preview/old.pl. On July 2003 there were 76 such predictions. This section of the PDB-Preview database forms the basis for the large-scale assessment experiment PDB-CAFASP (Fischer *et al*., 2003) (see also http://www.cs.bgu.ac.il/∼dfischer/CAFASP).

The PDB-Preview high-scoring FR predictions (obtained by 3D-Jury at scores >50) thus provide biologists with relatively accurate 3D models for 'on-hold' PDB entries in a timely fashion, shortly after they are deposited in the PDB, and well before the experimental 3D structure is released. In addition, the resulting PDB-CAFASP analysis provides computational biologists with a continuous blind evaluation of their methods, thus effectively extending other benchmarking experiments such as LiveBench and CAFASP.

ACKNOWLEDGEMENTS

The PDB-Preview database is only possible with the support of structure prediction servers. We would like to thank again all developers of the methods and services. This work was supported by the SPUB-M and 6FP grant MiFriend (QLRT-2001-02884) to L.R.

REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F., 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.

Bujnicki,J.M., Elofsson,A., Fischer,D. and Rychlewski,L. (2001) Structure prediction meta server. *Bioinformatics*, **17**, 750–751.

Fischer,D. and Rychlewski,L. (2003) The 2002 olympic games of protein structure prediction. *Protein Eng.*, **16**, 157–160.

Fischer,D., Rychlewski,L., Dunbrack,R.L.,Jr., Oritz,A.R. and Elofsson,A. (2003) CAFASP3, the third critical assessment of fully automated structure prediction methods. *Proteins*, **53** (Suppl. 6), 503–516.

Ginalski,K., Elofsson,A., Fischer,D. and Rychlewski,L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, **19**, 1015–1018.

Kinch,L.N., Wrabl,J.O., Krishna,S.S., Majumdar,I., Sadreyev,R.I., Qi,Y., Pei,J., Cheng,H. and Grishin,N.V. (2003) CASP5 assessment of fold recognition target predictions. *Proteins*, **53** (Suppl. 6), 395–409.

Koh,I.Y., Eyrich,V.A., Marti-Renom,M.A., Przybylski,D., Madhusudhan,M.S., Eswar,N., Grana,O., Pazos,F., Valencia,A., Sali,A. and Rost,B. (2003) EVA: evaluation of protein structure prediction servers. *Nucleic Acids Res.*, **31**, 3311–3315.

Li,W., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.

Rychlewski,L., Fischer,D. and Elofsson,A. (2003) LiveBench-6: large-scale automated evaluation of protein structure prediction servers. *Proteins*, **53** (Suppl. 6), 542–547.

Sussman,J.L., Lin,D., Jiang,J., Manning,N.O., Prilusky,J., Ritter,O. and Abola,E.E. (1998) Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr. D Biol. Crystallogr.*, **54**, 1078–1084.