BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 8 2001
Pages 750–751

Structure prediction meta server

Janusz M. Bujnicki[1], Arne Elofsson[2], Daniel Fischer[3] and Leszek Rychlewski[4,*]

[1]Bioinformatics Laboratory, International Institute of Molecular and Cell Biology (IIMCB), Warsaw, Poland, [2]Stockholm Bioinformatics Center, Stockholm University, 10691 Stockholm, Sweden, [3]Bioinformatics, Department of Computer Science, Ben Gurion University, Beer-Sheva 84105, Israel and [4]Bioinformatics Laboratory, BioInfoBank Institute, ul. Limanowskiego 24A, 60-744 Poznan, Poland

Received on March 5, 2001; revised and accepted on April 12, 2001

ABSTRACT

Summary: The Structure Prediction Meta Server offers a convenient way for biologists to utilize various high quality structure prediction servers available worldwide. The meta server translates the results obtained from remote services into uniform format, which are consequently used to request a jury prediction from a remote consensus server Pcons.

Availability: The structure prediction meta server is freely available at http://BioInfo.PL/meta/, some remote servers have however restrictions for non-academic users, which are respected by the meta server.

Contact: leszek@bioinfo.pl

Supplementary information: Results of several sessions of the CAFASP and LiveBench programs for assessment of performance of fold-recognition servers carried out via the meta server are available at http://BioInfo.PL/services.html.

The Structure Prediction Meta Server is aimed to provide fast and convenient assignment of three-dimensional structure for the query protein based only on the provided amino acid sequence. The meta server is a framework for communication between several providers of structure prediction services and offers a convenient interface for the user. The user is asked to provide the amino acid sequence of the query protein, the name of the query protein or a reference name for the prediction job, and the e-mail address. The e-mail address is used only for notification about errors during the execution of the job. The query sequence and the reference name are placed in a process queue coupled to an SQL database engine. The meta server accepts only sequences, which have not been submitted before. The possibility to update the predictions for sequences, which were submitted previously will be introduced to the meta server in the future. The SQL database offers the possibility to find any previous jobs processed by the meta server using 'regular expressions' addressing fields like e-mail, job name and the host name, from which the job was initiated. The database currently holds about 3000 predictions. In the future a purging process will delete very old jobs. The meta server is only a set of programs aimed to process and manage biological data, while the predictive power of the service comes from remote prediction providers that collaborate with the central meta server.

The meta server utilizes various, carefully evaluated services available from the community of the developers of structure prediction methods. The remote servers provide secondary (local) structure predictions as well as tertiary structure predictions and solvent accessibility. The local structure predictions are obtained from PsiPred (McGuffin *et al.*, 2000), Target99 (Karplus *et al.*, 1998) and a local-structure prediction meta server Jpred2 (Cuff and Barton, 2000). The current list of remote tertiary structure prediction (fold recognition) servers includes: FFAS (Rychlewski *et al.*, 2000), 3D-PSSM (Kelley *et al.*, 2000), GenTHREADER and mGenTHREADER (Jones, 1999), INBGU (Fischer, 2000), Sam-T99 (Karplus *et al.*, 1998), and FUGUE (Shi *et al.*, 2000). In addition, local installations of PDB-Blast (Bujnicki *et al.*, 2001), FFAS and 123D+ (Alexandrov *et al.*, 1996) are used. Each server has its own process queuing system managed by the meta server. This ensures very restrictive utilization of the computer power offered by the remote servers. A new request is submitted to a remote server only after the last prediction has been completed and the results have arrived. Requests that have been left unanswered for longer than 24 h are marked with 'error' and the next request is sent out.

All results of fold recognition servers are translated into uniform formats. The information extracted from the raw output of the servers includes the names (database codes) of the proteins with known structures (Bernstein *et al.*, 1977), which were matched to the query, the alignments of their amino acid sequence to the sequence of the

*To whom correspondence should be addressed.

query and the scores of the match which correspond to reliability estimations specific for every server. The data is used to produce the uniform output of the meta server which links the PDB codes with appropriate pages in the PDB database and extracts the structural classification for those proteins from the SCOP (Lo *et al.*, 2000) and FFSP (Holm and Sander, 1997) databases. If the PDB codes are absent from the structural databases, the sequence of the hits is compared to the classified proteins using BLAST (Altschul *et al.*, 1997) to retrieve the classification of the most similar protein. The secondary structure assignments for all PDB hits are taken from the FSSP database and the aligned protein sequence of the PDB proteins are colored accordingly (red for helices and blue for strands). The meta server also generates output for all alignments in standard formats like PDB, CASP or PIR.

The meta server is coupled to a consensus server (Pcons), which takes as queries the PDB-formatted output of several servers, and sorts all models based on their likelihood using internal judgement criteria (Lundström *et al.*, 2001). The meta server creates such input files after a translation of prediction results into PDB format, which are consequently submitted to Pcons. Pcons uses a neural network architecture, which was trained on a large set of results obtained from most coupled services. The current version of Pcons takes into account the models generated by PDB-Blast, FFAS, 3D-PSSM, INBGU, SamT99, Fugue, GenTHREADER and mGenTHREADER.

All tertiary structure prediction servers are continuously evaluated by a related Benchmarking Program LiveBench (Bujnicki *et al.*, 2001). LiveBench uses the meta server infrastructure for weekly submission of test predictions requests to the coupled servers. The test cases are selected from the new proteins deposited at the PDB, which don't show any obvious similarity to other proteins with known structure. LiveBench thus offers an *a posteriori* estimation of the reliability of the servers and mapping of their specific reliability scores on the expected probability of wrong structure assignment. Two benchmarking sessions were already completed and the results are available on the web pages of the program (http://BioInfo.PL/LiveBench/). Those results were also provided for training of the Pcons server and the evaluation of the accuracy of the consensus procedure was conducted. The results indicate that Pcons combines the sensitivity of the best fold recognition servers with very high specificity. Utilization of various high-performance structure prediction servers and an efficient jury evaluation procedure implemented in the meta server represents the most accurate automated structure prediction protocol currently available.

Another similar meta server is available at: http://www.infobiosud.univ-montp1.fr/SERVEUR/HTML_BIO/bioserver.html

REFERENCES

Alexandrov,N.N., Nussinov,R. and Zimmer,R.M. (1996) Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. In Hunter,L. and Klein,T. (eds), *Biocomputing: Proceedings of the 1996 Pacific Symposium*. World Scientific, Singapore, pp. 53–72.

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.

Bernstein,F.C., Koetzle,T.F., Williams,G.J., Meyer,E,J., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.

Bujnicki,J.M., Elofsson,A., Fischer,D. and Rychlewski,L. (2001) LiveBench—1: continuous benchmarking of protein structure prediction servers. *Protein Sci.*, **10**, 352–361.

Cuff,J.A. and Barton,G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502–511.

Fischer,D. (2000) Hybrid fold recognition: combining sequence derived properties with evolutionary information. *Pac. Symp. Biocomput.*, 119–130.

Holm,L. and Sander,C. (1997) Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.*, **25**, 231–234.

Jones,D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, **287**, 797–815.

Karplus,K., Barrett,C. and Hughey,R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.

Kelley,L.A., McCallum,C.M. and Sternberg,M.J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, **299**, 501–522.

Lo,C.L., Ailey,B., Hubbard,T.J., Brenner,S.E., Murzin,A.G. and Chothia,C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **28**, 257–259.

Lundström,D., Rychlewski,L., Bujnicki,J.M. and Elofsson,A. (2001) Pcons: a neural network based consensus predictor that improves fold recognition. *Protein Sci.*, submitted.

McGuffin,L.J., Bryson,K. and Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.

Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.

Shi,J., Blundell,T.L. and Mizuguchi,K. (2000) FUGUE profile library search against HOMSTRAD http://www-cryst.bioc.cam.ac.uk/~fugue/.