Nucleic Acids Research, 2003, Vol. 31, No. 13 **3305–3307**
DOI: 10.1093/nar/gkg557

GeneSilico protein structure prediction meta-server

Michal A. Kurowski and Janusz M. Bujnicki*

Bioinformatics Laboratory, International Institute of Molecular and Cell Biology, Warsaw, Poland

Received February 24, 2003; Revised and Accepted April 1, 2003

ABSTRACT

Rigorous assessments of protein structure prediction have demonstrated that fold recognition methods can identify remote similarities between proteins when standard sequence search methods fail. It has been shown that the accuracy of predictions is improved when refined multiple sequence alignments are used instead of single sequences and if different methods are combined to generate a consensus model. There are several meta-servers available that integrate protein structure predictions performed by various methods, but they do not allow for submission of user-defined multiple sequence alignments and they seldom offer confidentiality of the results. We developed a novel WWW gateway for protein structure prediction, which combines the useful features of other meta-servers available, but with much greater flexibility of the input. The user may submit an amino acid sequence or a multiple sequence alignment to a set of methods for primary, secondary and tertiary structure prediction. Fold-recognition results (target-template alignments) are converted into full-atom 3D models and the quality of these models is uniformly assessed. A consensus between different FR methods is also inferred. The results are conveniently presented on-line on a single web page over a secure, password-protected connection. The GeneSilico protein structure prediction meta-server is freely available for academic users at http://genesilico.pl/meta.

INTRODUCTION

The value of a protein's three-dimensional (3D) structure in connection with its molecular function is enormous because it provides a solid framework for planning experiments and for the interpretation of their results. Since experimental structure determination is very expensive and is not always successful, theoretical structure prediction became an important area of modern biology. There are several initiatives undertaken by the protein structure prediction community to provide an assessment of the capabilities and limitations of current methods for protein structure predictions: CASP (1), CAFASP (2), Livebench (3) and EVA (4). A major finding from the latest assessments is that better structure predictions can be obtained by combining the results produced using several different methods because they have different strengths and weaknesses. The CASP4 experiment showed that the group named *CAFASP-consensus*, which filed predictions extracted from a number of automated servers, performed considerably better than any individual server and better than all but six human predictors (5). In the last CASP5 experiment, the success of various 'meta-servers' and groups that used them to judiciously combine results obtained by several different methods was evident in all 3D structure modeling categories—from Comparative Modeling (CM), to Fold Recognition (FR), to Novel Fold (NF) prediction, as well as in the secondary structure (SS) prediction category (http://predictioncenter.llnl.gov/casp5/).

As reported by others (6,7) and in our hands as well, the use of manually refined multiple sequence alignments (MSA) as structure prediction queries gives significant improvement in the model quality (agreement with the real structure) over predictions based on single sequences. Most of the individual structure prediction methods (SS prediction as well as FR) allow the user to submit his/her own alignment or provide a BLAST or PSI-BLAST (8) utility to automatically build a MSA. The quality of MSA obtained by automatic methods is usually acceptable but user-defined alignments become clearly superior if the query protein has little or no close homologs in the sequence database used by default by the FR server. Moreover, the divergence of the query sequence (for instance the presence of very long loops) often leads to significant errors in the automatically-generated alignments. During the recent CASP4 and CASP5 experiments, in many cases we were able to obtain confident predictions of the correct fold only when we submitted a refined MSA, which repeatedly included additional sequences obtained from unfinished genomes or the EST databases. Such sequences are not available in the default databases and sometimes allow to increase the size of MSA more than 5-fold—this is critical when one compares the evolutionary information contained in the automatic alignment of 2–5 sequences and in the user-defined MSA of 10–25 sequences. Accordingly, when we submitted single sequences for automatic MSA building for such targets, prediction results often became ambiguous (data not shown). Submission of manually refined MSAs to FR servers allowed us to achieve high rankings—consistently within the best groups in both CM and FR categories

*To whom correspondence should be addressed. Tel: +48 226685384; Fax: +48 226685288; Email: iamb@genesilico.pl [i.e. CASP5 (9) and the assessment summary for BioInfo.PL in CASP4 (5,10)].

The existing meta-servers (for instance those available at http://bioinfo.pl or http://bioserv.infobiosud.univ-montp1.fr) provide a convenient interface for submission of prediction queries to multiple methods, unified presentation of their results and inference of a rational consensus. However, they do not allow the user to submit a MSA to the FR servers, which in our opinion is a critical issue in 3D structure prediction. Besides, the issue of confidentiality of results is not always addressed—for instance the BioInfo metaserver (11) makes a list of all prediction queries and the addresses of the computers from which the queries were submitted and also makes the prediction results freely available to everybody. This may be strongly discouraging for those users who don't want the prediction query (which may be a novel protein) or the results of the analysis to be revealed to any third party.

METHODS

Our aim was to create a convenient, secure and simple on-line structure prediction service for users who prefer not to sacrifice the quality for speed by unreservedly relying on automatic database searches, but choose to submit manually refined sequence alignments in order to obtain potentially more accurate predictions. Hence, we developed a novel WWW 'meta-server' as a gateway to several protein structure prediction methods, which addresses the two aforementioned key issues (i.e. MSA submission and data confidentiality). The GeneSilico server facilitates the access to several structure prediction methods through a single, secure and user-friendly WWW interface. Its architecture allows easy web scripting, which greatly facilitates automated submission and retrieval of data by clients (user-agents) based on the xml-rpc serialization. Conforming to the object-oriented programming standards, each page is actually an object that can be serialized and/or treated as a method. Our server is freely available at the URL http://genesilico.pl/meta for academic users who sign a license agreement. Some of the components may be unavailable for commercial users who are nevertheless welcome to contact us in order to obtain a separate limited license.

The user has several options for submission of the prediction query. Our server accepts both single sequences and alignments. If a single sequence is submitted, each method generates its own MSA, as in the other meta-servers. If a MSA is submitted, the user can choose between submission of a full-length query or limiting the analysis to regions with less than 30% gaps in the alignment. The second option allows to remove highly divergent loops, which often cause problems when matching the core structures of the template and the target. The user-defined MSA is submitted to those FR servers which allow this format of submission (see the http://genesilico.pl/meta web site for details). For submission to those servers which accept only single sequences and always build their own MSA, the user-defined MSA is converted into a 'consensus sequence'—again, the user has the freedom to choose between several alternative methods for consensus generation. We have tested the value of MSA submission during the recent CASP5 experiment—the quality of the target-template alignments we obtained from our meta-server was exceptional, which greatly helped us to 'win' the homology modeling contest (9).

The currently installed components of the GeneSilico meta-server include:

1. The HMMPFAM tool for the primary structure analysis (identification of PFAM domains) (12).
2. Secondary structure prediction methods PSIPRED (13), SAM-T02 (14) and PROF (15).
3. Methods for identification of potential transmembrane helices (to our knowledge, this type of method is not available via the other FR meta-servers): MEMSAT2 (16), TMHMM (17), TMPRED (18).
4. A local PDB-BLAST filter for identification of closely related sequences of known structures in PDB [prediction is halted and the user is notified if E-value $<10^{-3}$ is reported in the PSI-BLAST (8) search against the PDB database (19)].
5. 3D structure prediction core, which comprizes the best-performing FR methods currently available, according to the CAFASP and LIVEBENCH experiments: RAPTOR (20), 3DPSSM (21), FUGUE (22), mGENTHREADER (23), FFAS (24), SAM-T02 (14) and BIOINBGU (25) (at the time of the writing, the BIOINBGU server was temporarily unavailable).
6. The FR results returned by the above-mentioned servers are analyzed using the consensus server PCONS (26), which does not produce *de novo* results, but serves as a ranking system for selection of the potentially best models among those reported by the 'original' methods. PCONS was shown to perform much better than any individual FR server in the recent CASP (1), CAFASP (2) and Livebench (3) analyses. Among the available FR methods, the present version PCONS2 is able to analyze only the results of PDB-BLAST, 3DPSSM, FUGUE, mGENTHREADER and BIOINBGU. A new, updated version is planned for the near future which will also include RAPTOR, FFAS3 and SAMT-02.
7. Based on the FR results (target-template alignments), preliminary 3D models of the query structure are built using SCWRL (27), based on the backbone of the template. These crude models lack the features corresponding to gaps in the FR alignment (for instance insertions in the target, absent from the template), but the structure of the hydrophobic core is usually inferred well enough to perform 3D structure evaluation using VERIFY3D (28). Thereby, all FR alignments obtained from different servers undergo unified assessment by energetic criteria implemented in VERIFY3D in addition to the ranking criterion offered by the PCONS server.

The GeneSilico meta-server is continuously upgraded and enhanced with new tools. We hope that it will be as useful for the wide community as was for us in the CASP5 experiment, as well as in our daily work on protein sequence analysis and structure prediction.

ACKNOWLEDGEMENTS

We are grateful to developers of the protein fold recognition servers who kindly agreed to have them included in our meta-server, in particular Drs Arne Elofsson, Daniel Fischer, Adam Godzik, David Jones, Kevin Karplus, Lawrence Kelley, Kenji Mizuguchi and Jinbo Xu. J.M.B. is a fellow of the Foundation for Polish Science. This work was supported by the EMBO and HHMI Young Investigator Programme award to J.M.B. and by KBN (grants 3P04A 011 24 to J.M.B. and 3P05A 020 24 to M.A.K.).

REFERENCES

1. Moult,J., Fidelis,K., Zemla,A. and Hubbard,T. (2001) Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins*, **45** (Suppl. 5), 2–7.
2. Fischer,D., Elofsson,A., Rychlewski,L., Pazos,F., Valencia,A., Rost,B., Ortiz,A.R. and Dunbrack,R.L.,Jr (2001) CAFASP2: the second critical assessment of fully automated structure prediction methods. *Proteins*, **45** (Suppl. 5), 171–183.
3. Bujnicki,J.M., Elofsson,A., Fischer,D. and Rychlewski,L. (2001) LiveBench-2: large-scale automated evaluation of protein structure prediction servers. *Proteins*, **45**, 184–191.
4. Eyrich,V.A., Marti-Renom,M.A., Przybylski,D., Madhusudhan,M.S., Fiser,A., Pazos,F., Valencia,A., Sali,A. and Rost,B. (2001) EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics*, **17**, 1242–1243.
5. Sippl,M.J., Lackner,P., Domingues,F.S., Prlic,A., Malik,R., Andreeva,A. and Wiederstein,M. (2001) Assessment of the CASP4 fold recognition category. *Proteins*, **45** (Suppl. 5), 55–67.
6. Cuff,J.A. and Barton,G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502–511.
7. Williams,M.G., Shirai,H., Shi,J., Nagendra,H.G., Mueller,J., Mizuguchi,K., Miguel,R.N., Innis,C.A., Deane,C.M., Chen,L. *et al.* (2001) Sequence-structure homology recognition by iterative alignment refinement and comparative modeling. *Proteins*, **45** (Suppl. 5), 92–97.
8. 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.
9. Kosinski, J., Cymerman, I.A., Feder, M., Kurowski, M.A., Sasin, J.M. and Bujnicki,J.M. (2003) A 'Frankenstein's monster' approach to comparative modeling: merging the finest fragments of fold-recognition models and iterative model refinement aided by 3D structure evaluation. *Proteins*, in press.
10. Tramontano,A., Leplae,R. and Morea,V. (2001) Analysis and assessment of comparative modeling predictions in CASP4. *Proteins*, **45** (Suppl. 5), 22–38.
11. Bujnicki,J.M., Elofsson,A., Fischer,D. and Rychlewski,L. (2001) Structure prediction Meta Server. *Bioinformatics*, **17**, 750–751.
12. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
13. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
14. Karplus,K., Karchin,R., Barrett,C., Tu,S., Cline,M., Diekhans,M., Grate,L., Casper,J. and Hughey,R. (2001) What is the value added by human intervention in protein structure prediction? *Proteins*, **45** (Suppl. 5), 86–91.
15. Ouali,M. and King,R.D. (2000) Cascaded multiple classifiers for secondary structure prediction. *Protein Sci.*, **9**, 1162–1176.
16. Jones,D.T., Taylor,W.R. and Thornton,J.M. (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **33**, 3038–3049.
17. Sonnhammer,E.L., von Heijne,G. and Krogh,A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 175–182.
18. Hofmann,K. and Stoffel,W. (1993) TMbase—a database of membrane spanning proteins segments. *Biol.Chem.*, **374**, 166
19. Westbrook,J., Feng,Z., Chen,L., Yang,H. and Berman,H.M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **31**, 489–491.
20. Xu,J., Li,M., Lin,G., Kim,D. and Xu,Y. (2003) Protein structure prediction by linear programming. *Pac. Symp. Biocomput.*, 264–275.
21. 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.
22. Shi,J., Blundell,T.L. and Mizuguchi,K. (2001) Fugue: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.*, **310**, 243–257.
23. Jones,D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, **287**, 797–815.
24. 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.
25. Fischer,D. (2000) Hybrid fold recognition: combining sequence derived properties with evolutionary information. *Pac. Symp. Biocomput.*, 119–130.
26. Lundstrom,J., Rychlewski,L., Bujnicki,J.M. and Elofsson,A. (2001) Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci.*, **10**, 2354–2362.
27. Dunbrack,R.L. (1999) Comparative modeling of CASP3 targets using PSI-BLAST and SCWRL. *Proteins* (Suppl. 3), 81–87.
28. Luthy,R., Bowie,J.U. and Eisenberg,D. (1992) Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.