

META-PP: single interface to crucial prediction servers

Volker A. Eyrich^{1,*} and Burkhard Rost^{1,2,3}

¹CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, ²Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St Nicholas Avenue and ³North East Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, USA

Received February 15, 2003; Revised and Accepted April 8, 2003

ABSTRACT

The META-PP server (<http://cubic.bioc.columbia.edu/meta/>) simplifies access to a battery of public protein structure and function prediction servers by providing a common and stable web-based interface. The goal is to make these powerful and increasingly essential methods more readily available to non-expert users and the bioinformatics community at large. At present META-PP provides access to a selected set of high-quality servers in the areas of comparative modelling, threading/fold recognition, secondary structure prediction and more specialized fields like contact and function prediction.

INTRODUCTION

Pros and cons of making web servers easily available. Developers in computational biology and bioinformatics increasingly make their methods available to a broader audience of biologists through web-based interfaces. In fact commonly, even ‘bleeding-edge’ methods are available through the web before they are announced to the scientific community via traditional routes of publication. The resulting explosion in the number of public servers is both wonderful and problematic. The major advantages are that researchers can choose from a large number of prediction methods (and the underlying expertise and expert knowledge) that attempt to address a broad range of problems without having to maintain software on-site. Methods are readily available from a web browser and updates to the servers can be deployed by the administrators in a manner that is transparent to the end-user—note that the latter can also be a disadvantage for users who expect a certain amount of stability in the underlying methodologies. The disadvantages are that non-experts might be overwhelmed by the sheer number of available servers and might be unable to judge the quality of the resulting predictions since there is no widely accepted ‘quality control

process’ as for traditional peer-reviewed publications. There are many examples of methods that are sub-standard but have more attractive interfaces than do the corresponding state-of-the-art tools. Furthermore, methods and web sites undergo constant changes, some of these long-term users may never realize, and some of these may just lead to frustration since the servers no longer work as advertised.

CASP: a field establishes self-control. In 1994, John Moult initiated a unique concept: establish a procedure for critically assessing the quality of structure prediction methods (CASP) (1–4). In other words, a field began to establish its internal quality control mechanism that went beyond what is possible in traditional peer review. The bi-annual CASP experiments illustrate some of the pros and cons of having servers readily available (1). The number of servers for comparative modelling, fold recognition/threading and secondary structure prediction has increased from 12 in 1998 at CASP3/CAFASP1 when automatic servers were first analysed systematically (5) to >60 at the most recent CAFASP3 in 2002. Methods taking part in CAFASP often implement cutting-edge prediction approaches of a quality and reliability that has yet to be determined and it is often very difficult to understand the strengths and weaknesses of particular methods that are available. [Note: this problem is addressed by EVA, our server that automatically and continuously evaluates prediction servers (6–8).] Furthermore, the interfaces to these often cutting-edge methods are frequently such that only expert users patient enough for long navigation succeed in obtaining results for their queries. Clearly, a more streamlined and uniform method of accessing these services would be of use to both the casual and the expert user. To provide such a simple interface that allows a user to type in the sequence once and to obtain results from many servers was one motivation for developing META-PP. The other was to restrict the list of servers to those with sustained state-of-the-art performance that returned results reliably over long periods of time (for details: <http://cubic.bioc.columbia.edu/meta/criteria.html>). Other groups have made similar interfaces available to the bioinformatics

*To whom correspondence should be addressed. Tel: +1 2123053773; Fax: +1 2123057932; Email: volker@chem.columbia.edu

community and we encourage users to visit the following web sites:

- <http://bioinfo.pl/meta> and <http://genesilico.pl/meta>, two resources that focus mostly on fold recognition and threading servers;
- <http://searchlauncher.bcm.tmc.edu> and <http://workbench.sdsc.edu>, two sites that provide access to a wide variety of sequence search and sequence alignment tools, secondary structure prediction servers and gene feature search tools as well as more specialized resources.

Naturally, there is some overlap in the methods that the above resources provide access to and the servers that are included in META-PP. The major difference between META-PP and other meta-servers is the care taken by META-PP to restrict the list of services to sustained methods. This increases the reliability while reducing the mere number of available methods.

METHODS

Engine behind META-PP. As outlined above the META-PP server provides a simple streamlined interface to a wide range of prediction servers in computational biology/bioinformatics. Users access the server via a simple web interface (<http://cubic.bioc.columbia.edu/meta/>). Input is a one-letter code protein sequence in one of several common formats along with a short description of the protein (optional) and an email address. Users then select the sub-set of available servers they want to access. META-PP validates the input (email address and sequence format) and places the request into a processing queue. Requests are scheduled for processing immediately leading to very short turn-around times for the user in most cases—in particular since META-PP can process several predictions in parallel. During the processing of a prediction request META-PP assembles the raw data required for submission such as sequences and job options, connects to the remote server using the appropriate protocol and submits the request. Depending on the server META-PP might wait and receive actual output in real-time or simply wait for submission confirmation and then disconnect. In the case of failure, caused, for example, by intermittent outages at the remote site or by simple connectivity problems, META-PP reinserts the failed request back into its own processing queue and resubmits at a later time (for up to 24 h after which failed prediction requests are simply purged from the processing queue). Depending on the characteristics of the prediction server, users will receive results either from META-PP or directly from the original prediction server. At present, META-PP does not—in contrast to PredictProtein (9)—attempt to reformat the data obtained from remote servers in order to preserve the integrity of the results. An example for results returned by META-PP is available at: <http://cubic.bioc.columbia.edu/meta/help.html>.

Prediction categories. META-PP provides access to a larger number of prediction servers in several categories and not just the more popular ones like comparative modelling, fold recognition/threading and secondary structure prediction. Currently the list of servers that META-PP provides access to includes CPHmodels (10), ChloroP (11), DAS, Jpred (12,13), NetOglyc (14), NetPhos, NetPico, PSIPred (15,16), Phd (17,18),

Prof, SAM-T99 (19), SSPro (20,21), SignalP (22), Superfamily (23), Swiss-Model (24,25), TMHMM 2.0 (26) and TopPred2 (27) among others. Note that META-PP is very extensible and the list presented here will change over time as new methods become available or existing ones are being updated. It should be noted though that META-PP only includes prediction servers if their developers explicitly agree.

Statistics after 2.5 years of operation. META-PP went online in the summer of 2000 at Columbia University. Currently we receive an average of ~1000 requests per day and since its inception, META-PP has processed >500 000 prediction requests from well over 10 000 individual users. Note also that META-PP serves as the submission engine used in the EVA experiment (6,8).

CONCLUSIONS

META-PP provides streamlined access to a select subset of prediction servers to the bioinformatics community at large and in particular to non-expert users. It has proved to be a convenient and reliable means of navigating the large number of prediction servers currently available on the World Wide Web and we hope that it will continue to provide a useful service for researchers interested in the prediction of protein structure and function.

ACKNOWLEDGEMENTS

Thanks to Jinfeng Liu (Columbia) for computer assistance. The work of B.R. was supported by grants 1-P50-GM62413-01 and RO1-GM63029-01 from the National Institute of Health (NIH), and the grant DBI-0131168 from the National Science Foundation (NSF). Last, and by no means least, thanks to all those who deposit their experimental data in public databases, those who maintain these databases and all the developers who make their prediction methods available on the World Wide Web.

REFERENCES

1. Moul, J., Pedersen, J.T., Judson, R. and Fidelis, K. (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins*, **23**, ii–iv.
2. Moul, J., Hubbard, T., Bryant, S.H., Fidelis, K. and Pedersen, J.T. (1997) Critical assessment of methods of protein structure prediction (CASP): Round II. *Proteins*, **29** (Suppl. 1), 2–6.
3. Moul, 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.
4. Moul, J., Hubbard, T., Bryant, S.H., Fidelis, K. and Pedersen, J.T. (1999) Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins*, **37** (Suppl. 3), 2–6.
5. Fischer, D., Barret, C., Bryson, K., Elofsson, A., Godzik, A., Jones, D., Karplus, K.J., Kelley, L.A., MacCallum, R.M., Pawowski, K. *et al.* (1999) CAFASP-1: Critical assessment of fully automated structure prediction methods. *Proteins*, **37**, 209–217.
6. 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.

7. Marti-Renom, M.A., Madhusudhan, M.S., Fiser, A., Rost, B. and Sali, A. (2002) Reliability of assessment of protein structure prediction methods. *Structure*, **10**, 435–440.
8. Rost, B. and Eyrich, V.A. (2001) EVA: Large-scale analysis of secondary structure prediction. *Proteins*, **45**, 192–199.
9. Rost, B. and Liu, J. (2003) The PredictProtein server. *Nucleic Acids Res.*, **31**, 3300–3304.
10. Lund, O., Frimand, K., Gorodkin, J., Bohr, H., Bohr, J., Hansen, J. and Brunak, S. (1997) Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng.*, **10**, 1241–1248.
11. Emanuelsson, O., Nielsen, H. and Von Heijne, G. (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.*, **8**, 978–984.
12. Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, M. and Barton, G.J. (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics*, **14**, 892–893.
13. Cuff, J.A. and Barton, G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502–511.
14. Hansen, J.E., Lund, O., Tolstrup, N., Gooley, A.A., Williams, K.L. and Brunak, S. (1998) NetOglyc: Prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glyco. J.*, **15**, 115–130.
15. Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
16. McGuffin, L.J., Bryson, K. and Jones, D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.
17. Rost, B. and Sander, C. (1993) Prediction of protein secondary structure at better than 70 percent accuracy. *J. Mol. Biol.*, **232**, 584–599.
18. Rost, B. and Sander, C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, **19**, 55–72.
19. Karplus, K., Barrett, C. and Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.
20. Baldi, P., Brunak, S., Frasconi, P., Soda, G. and Pollastri, G. (1999) Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, **15**, 937–946.
21. Pollastri, G., Przybylski, D., Rost, B. and Baldi, P. (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, **47**, 228–235.
22. Nielsen, H., Brunak, S. and von Heijne, G. (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.*, **12**, 3–9.
23. Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.
24. Guex, N. and Peitsch, M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
25. Guex, N., Diemand, A. and Peitsch, M.C. (1999) Protein modelling for all. *Trends Biochem. Sci.*, **24**, 364–367.
26. Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
27. Claros, M.G. and Von Heijne, G. (1994) Toppred-II—an improved software for membrane-protein structure predictions. *Comp. Appl. Biosci.*, **10**, 685–686.