BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 12 2001
Pages 1242–1243

EVA: continuous automatic evaluation of protein structure prediction servers

Volker A. Eyrich[1], Marc A. Martí-Renom[2], Dariusz Przybylski[3], Mallur S. Madhusudhan[2], András Fiser[2], Florencio Pazos[4], Alfonso Valencia[4], Andrej Sali[2] and Burkhard Rost[3,*]

[1]Columbia University, Department of Chemistry, 3000 Broadway MC 3136, New York, NY 10027, USA, [2]The Rockefeller University, Laboratory of Molecular Biophysics, Pels Family Center for Biochemistry and Structural Biology, 1230 York Avenue, New York, NY 10021-6399, USA, [3]CUBIC Columbia University, Department of Biochemistry and Molecular Biophysics, 650 West 168th Street, New York, NY 10032, USA and [4]Protein Design Group, CNB-CSIC, Cantoblanco, Madrid 28049, Spain

Received on February 20, 2001; revised on May 28, 2001; accepted on July 4, 2001

ABSTRACT

Summary: Evaluation of protein structure prediction methods is difficult and time-consuming. Here, we describe EVA, a web server for assessing protein structure prediction methods, in an automated, continuous and large-scale fashion. Currently, EVA evaluates the performance of a variety of prediction methods available through the internet. Every week, the sequences of the latest experimentally determined protein structures are sent to prediction servers, results are collected, performance is evaluated, and a summary is published on the web. EVA has so far collected data for more than 3000 protein chains. These results may provide valuable insight to both developers and users of prediction methods.
Availability: http://cubic.bioc.columbia.edu/eva.
Contact: eva@cubic.bioc.columbia.edu

EVALUATING PREDICTIONS IS CRUCIAL

Correctly evaluating structure prediction is difficult. Developers of prediction methods in bioinformatics may significantly over-estimate their performance because of the following reasons. First, it is difficult and time-consuming to correctly separate data sets used for developing and testing. Second, estimates of performance of the different methods are often based on different data sets. This problem frequently originates from the rapid growth of the sequence and structure databases. Third, single numbers are usually not sufficient to describe the performance of a method. The lack of clarity is particularly unfortunate at a time when an increasing number of tools are made easily available through the internet and many of the users are not experts in the field of protein structure prediction.

*To whom correspondence should be addressed.

How well do experts predict protein structure? The CASP experiments attempt to address the problem of over-estimated performance (Zemla *et al.*, 2001). Although CASP resolves the bias resulting from using known protein structures as targets, it has limitations. (1) The methods are ranked by human assessors who have to evaluate thousands of predictions in 1–2 months (∼ 10 000 from 160 groups for CASP4; Zemla *et al.*, 2001). (2) Some aspects of the assessments are not statistically significant because they are based on few proteins. (3) The assessments cover only proteins determined in a period of about four months every two years. (4) Users cannot always reproduce CASP predictions, because programs or the required human expertise are not available. Effectively, CASP aims at assessing how well experts can predict structure.

How well do computers predict protein structure? CAFASP has recently extended CASP by testing automatic prediction servers on the CASP proteins (Fischer *et al.*, 1999). Although CAFASP aimed at evaluating programs rather than experts, it is still limited to a small number of test proteins (Zemla *et al.*, 2001). This limitation prompted us to create EVA, a large-scale and continuously running web server that automatically assesses protein structure prediction servers (http://cubic.bioc.columbia.edu/eva/doc/flow.html). The aims of EVA are: (1) Evaluate continuously and automatically blind predictions by all co-operating prediction servers. (2) Update the results on the web every week. (3) Enable developers, non-expert users, and reviewers to determine the performance of the tested prediction programs. (4) Compare prediction methods based on identical and sufficiently large data sets. Similar aims are also pursued by the LiveBench project (Rychlewski and Fischer, 2000). Although EVA continues to grow, most

of these objectives have already been realised. We will extend EVA in three additional ways: (i) test more servers, (ii) refine the evaluation of threading servers, and (iii) add alternative structure alignment methods for evaluation. EVA is already downloading target sequences from PDB prior to the release of their structures.

CURRENT IMPLEMENTATION OF EVA

Results in four prediction categories. Currently, EVA evaluates four different categories of structure prediction servers (see EVA home page for URLs and list of servers): comparative modelling (3), threading (6), secondary structure prediction (9), and inter-residue contact prediction (4). Brief explanations about the methods are on the EVA web site.

Results are updated every week. Every day, EVA downloads the newest protein structures from PDB (Berman *et al.*, 2000). The structures are added to a mySQL database, sequences are extracted for every protein chain, and sent to each server by META-PP (Eyrich and Rost, 2000). Predictions are collected and sent for evaluation to the EVA-satellites (comparative modelling: Rockefeller University, contacts: CNB Madrid, and all other: Columbia University). Depending on the category, the assessments are made available within hours to days. The central EVA site at CUBIC downloads all HTML pages produced by the satellites, and builds up the 'latest week' results that are then mirrored at the satellites (for a flowchart of EVA, see http://cubic.bioc.columbia.edu/eva/doc/flow.html).

Comparing: Identical data sets, major questions first! EVA compares methods based only on identical data sets. This approach is essential for reliably ranking methods. However, it reduces the number of available proteins since not all predictions are available for all servers. Another important feature of EVA is that it displays the results hierarchically, so that users get the 'big picture' first, followed by information at increasingly higher levels of detail upon request.

Methods are not ranked based on too few test proteins! Since prediction accuracy varies between proteins, published estimates for performance are averages over many proteins, with some standard deviation. We use this standard deviation to estimate the error of the average accuracy as a function of the test set size. This is justified, since different prediction methods typically have similar standard deviations. For example, when a method correctly predicts 75% of the residues in a set of 16 proteins with a standard deviation of 10%, a difference relative to another method $< 2.5\%$ ($\Delta Q = 10/\mathrm{sqrt}\,(16)$) is not significant. Thus, we cannot distinguish between 75% and 73% accuracy.

Resource with over 40 000 predictions. 2996 new protein structures have been added to PDB since EVA started in June 2000. The 2996 proteins were dissected into 3665 chains, 3130 (85%) of which were similar in sequence to known structures and 535 (15%) of which were new (less than 30% sequence identity over more than 100 residues). In comparative modelling, EVA evaluated over 6600 models with common subsets for 303 chains. In secondary structure prediction, EVA based its analysis on over 30 000 individual predictions; 127 chains were common to all methods, 348 to four methods. For both of these categories, EVA evaluated most of the existing servers in the field on the largest protein sets ever. Details about the evaluation are available on the EVA web site; details about the predictions will be published elsewhere.

Additional resources: PSI-BLAST alignments and sequence unique subset of PDB. EVA also maintains a number of additional data resources. One resource is a continuously updated list giving the largest subset of sequence-unique proteins in PDB (no protein in the set shares more than 33 identical residues over 100 residues aligned). This set now contains 2435 chains. Another resource contains over 5000 PSI-BLAST alignments for proteins added to PDB during the existence of EVA.

ACKNOWLEDGEMENTS

We are particularly grateful to Phil Bourne (UCSD) and Kevin Karplus (UCSC) for their support. We would also like to thank Arne Elofsson (Stockholm), Torsten Schwede, Nicolas Guex, and Manual Peitsch (all three form Glaxo, Geneva) for helpful discussions, and Nigel Brown (MRC, London) for his program MView. Last not least, we are grateful to the developers who permitted us to test their prediction servers. We apologise to all whose servers we evaluated that we had to remove their citations from this paper; they can be found at: http://cubic.bioc.columbia.edu/eva/doc/explain_methods.html.

REFERENCES

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Eyrich,V. and Rost,B. (2000) The META-PredictProtein server. WWW document (http://cubic.bioc.columbia.edu/predictprotein/submit_meta.html) CUBIC, Columbia University, Department of Biochemistry & Molecular Biophysics.

Fischer,D., Barret,C., Bryson,K., Elofsson,A., Godzik,A., Jones,D., Karplus,K.J., Kelley,L.A., MacCallum,R.M., Pawowski,K., Rost,B., Rychlewski,L. and Sternberg,M. (1999) CAFASP-1: critical assessment of fully automated structure prediction methods. *Proteins*, **3** (Suppl.), 209–217.

Rychlewski,L. and Fischer,D. (2000) LiveBench: continuous benchmarking of prediction servers. WWW document (http://BioInfo.PL/LiveBench/) http://BioInfo.PL/LiveBench/, IIMCB Warsaw.

Zemla,A., Venclovas,C. and Fidelis,K. (2001) Protein structure prediction center. http://PredictionCenter.llnl.gov/, Lawrence Livermore National Laboratory.