

EVA: continuous automatic evaluation of protein structure prediction servers

Volker A. Eyrich¹, Marc A. Martí-Renom², Dariusz Przybylski³, Mallur S. Madhusudhan², András Fiser², Florencio Pazos⁴, Alfonso Valencia⁴, Andrej Sali² & Burkhard Rost³

¹ Columbia Univ., Dept. of Chemistry, 3000 Broadway MC 3136, New York, NY 10027, USA

² The Rockefeller Univ., Lab. of Molecular Biophysics, Pels Family Center for Biochemistry and Structural Biology, 1230 York Avenue, New York, NY 10021-6399, USA

³ CUBIC Columbia Univ, Dept. of Biochemistry and Molecular Biophysics, 650 West 168th Street, New York, N.Y. 10032, USA

⁴ Protein Design Group, CNB-CSIC, Cantoblanco, Madrid 28049, Spain

Corresponding author: B Rost, rost@columbia.edu, Tel +1-212-305-3773, Fax +1-212-305-7932
Running title: EVA: evaluation of prediction servers
Document type: Application note
Statistics: Abstract 153 words; Text 1129 words; 5 References

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: All EVA web pages are accessible at {<http://cubic.bioc.columbia.edu/eva>}. Pages are mirrored at {<http://pipe.rockefeller.edu/~eva>} and {<http://pdg.cnb.uam.es/eva>}. Dynamic pages that are built upon a user query are restricted to the mirror-site that generated them. However, most pages are static and the entire HTML directory tree can be downloaded by ftp.

Contact: eva@cubic.bioc.columbia.edu

Key words: protein structure prediction, automatic evaluation, CASP, comparative modelling, secondary structure, inter-residue contacts

NOTE to the printer: Please maintain the italic typesetting for the first sentence of each paragraph instead of sub-headings.

Text

Evaluating structure prediction is an important objective

Correctly evaluating structure prediction methods is a hard problem. 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.

How well do experts predict protein structure? An attempt to address the problem of over-estimated performance has been made by the CASP experiments (Zemla, et al., 2001). Although CASP resolves the bias resulting from using known protein structures as prediction targets, it has at least four limitations. (1) The methods are ranked by human assessors who usually have one to two months to evaluate thousands of predictions (approximately 10,000 from 160 groups for CASP4 (Zemla, et al., 2001)). (2) Many aspects of the assessments are not statistically significant because they are based on a small number of proteins (e.g., 14 for comparative modelling at CASP4). (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 computer programs or the required human expertise are often 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. EVA is already downloading target sequences from PDB prior to the release of their structures. 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.

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 evaluated): comparative modelling (3 servers), threading (6 servers), secondary structure prediction (9 methods), and inter-residue contact prediction (4 methods). Brief explanations about the methods are available 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 are sent to each prediction server by META-PP (Eyrich and Rost, 2000). Predictions are collected and sent for evaluation to the EVA-satellites: to Rockefeller University for comparative modelling, to CNB Madrid for contact predictions, and to CUBIC at Columbia University for all other predictions. 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 Rockefeller University and at the CNB (for a flowchart of EVA, see <http://cubic.bioc.columbia.edu/eva/doc/flow.html>).

Comparing methods: identical data sets, major questions first! EVA compares different methods based only on the same set of target proteins. This approach is essential for reliably ranking methods. However, it reduces the number of proteins on which the methods can be evaluated 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! For example, because the accuracy of secondary structure prediction varies among different proteins, published estimates of prediction accuracy are typically averages over many test sequences, with standard deviations usually above ten percentage points. We use this standard deviation to estimate the error of the average accuracy as a function of the test set size. The observation that different prediction methods typically have similar standard deviations provides a necessary justification for this approach. For example, when a method correctly predicts 75% of the residues in a test set of 16 proteins with a

standard deviation of 10%, a difference relative to another method that is smaller than 2.5% (*ie*, $Q = 10/\sqrt{16}$) is not significant. Thus, we cannot distinguish between two methods that predict correctly 75% and 73% of all residues, respectively.

After more than one year of testing: a 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 had sequence similarity to previously known chains and 535 (15%) of which had no significant sequence similarity to known structures (less than 30% sequence identity over more than 100 residues aligned). In comparative modelling, EVA evaluated more than 6600 models with common subsets for 303 chains. In secondary structure prediction, EVA based its analysis on a total of over 30,000 individual predictions; common subsets comprised from 127 (all methods) to 348 (four methods) chains. 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. In addition to the evaluation of structure prediction, 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 set share more than 33 identical residues over 100 residues aligned). This set now contains 2435 chains. Another resource comprises currently over 5000 PSI-BLAST alignments for proteins added to PDB while EVA is running (both in ASCII and HTML).

Acknowledgements

We are particularly grateful to Phil Bourne (UCSD) and Kevin Karplus (UCSC) for their support. We also like to thank Arne Elofsson (Stockholm), Torsten Schwede, Nicolas Guex, and Manual Peitsch (all three from Glaxo, Geneva) for helpful discussions, and Nigel Brown (MRC London) for his program MView. The contact prediction and evaluation servers are maintained at the Complutense supercomputer centre (Madrid). Last not least, we are grateful to the developers who permitted us to test their prediction servers: Pierre Baldi (Irvine), Phil Bourne (UCSD), Søren Brunak (Copenhagen), James Cuff (London), Piero Fariselli (Bologna), Mitsuo Iwate (Kitasoto), Ross King (Aberystwyth), Ole Lund (Copenhagen), Jarek Meller (Ithaca), David Jones (Uxbridge), Kevin Karplus (UCSC), Osvaldo Olmea (Madrid), Gianluca Pollastri (Irvine), Gajendra Raghava (Chandigarh), Kristoffer Rapacki (Copenhagen), Torsten Schwede (Geneva). 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. *Nucl. 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, Dept. 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*, Suppl 3, 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.