Protein Engineering vol.16 no.3 pp.157–160, 2003
DOI: 10.1093/proeng/gzg022

COMMUNICATION

The 2002 Olympic Games of Protein Structure Prediction

Daniel Fischer[1,2] and Leszek Rychlewski[3]

[1]Bioinformatics, Department of Computer Science, Ben Gurion University, Beer-Sheva 84015, Israel and [3]BioInfoBank Institute, Limanowskiego 24A/16, 60-744 Poznan, Poland

[2]To whom correspondence should be addressed.
E-mail: dfischer@cs.bgu.ac.il

The summer of every even year is considered by the protein structure prediction community as the Olympic Games season, because in addition to a number of continuous benchmarking experiments such as LiveBench, much effort is invested in the blind prediction experiments CASP and CAFASP. Here we report the major advances registered in the field since the last Games of 2000, as measured by the recently completed LiveBench-4 experiment. These results provide a timely measure of the capabilities of current methods and of their expected performance in the upcoming CASP-5 and CAFASP-3 experiments. We also describe the initiation of the two new, community-wide experiments, PDB-CAFASP and MR-CAFASP. These new experiments extend the scope of previous efforts and may have important implications for structural genomics.

Keywords: CAFASP/CASP/LiveBench/protein structure prediction/structural genomics

Introduction

One of the challenges of the post-genomic era is to assign three-dimensional (3D) structures computationally to the proteins encoded in genome sequences (Fischer and Eisenberg, 1997; Abbott, 2001; Fischer *et al.*, 2001a). As a result of the various sequencing and structural genomics projects, structure prediction methods are playing an increasingly critical role in translating the information on the relatively small subset of proteins whose structures will be solved into accurate models for all proteins (Baker and Sali, 2001, Fischer *et al.*, 2001a). To understand the capabilities and limitations of current methods, a number of assessment experiments have been developed. In this paper we describe what we have learned from recent experiments, focusing in the sub-area of fold recognition (FR); for reviews, see elsewhere (Fischer and Eisenberg, 1999; Jones, 1999; Kelley *et al.*, 1999; Fischer, 2000; Rychlewski *et al.*, 2000; Shi *et al.*, 2001). FR methods use the structural information of solved proteins to model the structure of those proteins that share no significant sequence similarity to any of the proteins of known structure.

We first briefly describe the major prediction experiments and the main lessons learned from the 2000 session. In the CASP (CASP4, 2000) blind prediction experiment, a few dozen proteins of known sequence but unknown structure are used as prediction targets. Human predictors file their models before the experimental structures are determined. When the structures become available, the predictions are evaluated by expert human assessors. The Fully Automated version of CASP is the CAFASP experiment (Fischer *et al.*, 2001b). In CAFASP, the same prediction targets are used, but only the predictions of automated servers are considered (i.e. there is no human intervention in the prediction process). In addition, in CAFASP the evaluation of the predictions is carried out by fully automated evaluation programs, rather than by human assessors as in CASP. Thus, the evaluation results are completely objective and reproducible. The Fully Automated CAFASP experiments are valuable not only within the community of computational biologists, but also to biologists; what biologists want to know is which program they should use for their prediction targets and not which group was able to produce the best predictions at CASP. With the advent of genome sequencing projects and with the worldwide structure genomics initiatives, the need for fully automated structure prediction has become evident.

The large-scale version of CAFASP is the LiveBench (LB) experiment. LB assesses the servers using a relatively large number of prediction targets compiled every week from newly released protein structures. Another advantage of LB is that it runs continuously and not only once every two years. Thus, it can provide a snapshot of the predicting capabilities of the servers approximately every half year. Another related, large-scale evaluation project, not discussed here, is EVA (Eyrich *et al.*, 2001), which mainly includes the evaluation of automated homology modeling, secondary structure and contact prediction methods. CASP5, CAFASP3 and LB-6 are currently under way and their results were to be announced at the Asilomar meeting in December, 2002. Over 150 predicting groups worldwide have registered in CASP5; in CAFASP3 there are over 70 automated servers of which 30 are also participating in LB-6.

What have we learned from previous experiments?

Because the results of CASP-4, CAFASP-2 and LB-2 have been published (CASP4, 2000), we only briefly summarize some of the main findings, focusing on the FR servers.

Probably the most widely agreed upon conclusion is that the evaluation of the accuracy of predicted 3D models is very difficult (Siew and Fischer, 2001; Cristobal *et al.*, 2001). While most predictors agree on what could be considered an excellent or completely wrong prediction, there is much controversy on how to assess and grant credit to those that are only partially correct. The problem of delineating an exact ranking of the predictors arises because many groups file excellent predictions for the ‘easy’ prediction targets and most groups file completely wrong predictions for the hardest ones. Thus, the differences among most groups are due mainly to those targets in the middle, for which the assessment problem is most severe. In addition, if the test set is not very large, then slight fluctuations can have a significant effect in the final ranking.

Despite these difficulties, progress in evaluation methods has been achieved and it seems that it is possible to distinguish a group of best performing servers from the others. Taken

Table I. LiveBench-4 sensitivity results

Sensitivity range (%)	Names of servers							
49–58	ORFs*	SHGU*	3DPS	INBG	FUG2*	FFAS	MGTH	FUG3*
41–43	GETH	ORFb*	FUGU	SFAM*	ST99			
34	PDBb							

The sensitivity indicates the percentage of correctly predicted targets (out of 108), based on MaxSub's evaluation criteria. The performance of the LB-4 meta-servers is described in the text.
*New servers.

together, the results of previous CASP, CAFASP and LB experiments seem roughly to agree on a group of five top-performing servers (referred to here with their four-character code used in LB): 3DPS (3D-PSSM) (Kelley *et al.*, 1999), MGTH (Mgenthreader) (Jones, 1999), INBG (INBGU) (Fischer, 2000), FFAS (FFAS) (Rychlewski *et al.*, 2000) and FUGU (FUGUE) (Shi *et al.*, 2001), with ST99 (SAMT99) (Karplus *et al.*, 1998) following closely.

Another important lesson from the recent experiments has been that the best human predictors still outperform the automated methods. Understanding and analyzing the aspects of human expertise that lead to a better human performance is important to allow their future incorporation into automated programs; this is and will continue to be one of the major challenges for bioinformaticians (Siew and Fischer, 2001). Nevertheless, CASP4 and CAFASP2 demonstrated that the automated methods perform surprisingly well. Out of over 100 human groups that participated in CASP4, only 11 human groups were ranked by the CASP4 assessor above the best of the servers, 3DPS. Furthermore, at rank 7 was the semi-automated method named CAFASP-CONSENSUS, that filed predictions using the CAFASP results of the servers. This demonstrated that the use of the automated predictions of a number of servers could result in improved performance and may probably be considered one of the most valuable lessons from CASP4.

Owing to the large-scale nature of the LB experiments, some quantification of the servers' performance is possible. The LB-2 results indicated that the best servers are able to produce correct fold assignments for between one-third and one-half of all newly released structures that show no sequence similarity to other proteins of known structure. Another valuable quantitative measure computed in LB is the specificity performance. Knowing the specificity of any prediction program is essential for its wider applicability. LB-2 demonstrated that the servers' specificities need to be improved.

One of the most significant results of LB-2 was the superior performance of the automated meta-predictor pcons (Lundstrom *et al.*, 2001). Pcons automates some of the procedures used by the CAFASP-CONSENSUS in CASP4 and selects one of the models from those produced by a number of servers. As expected, and consistently with the CASP4/CAFASP2 results, pcons generally outperformed the individual servers.

For further details on the LB-2 and CAFASP2 see (Bujnicki *et al.*, 2001; Fischer *et al.*, 2001b) and the comprehensive tables available at the corresponding web pages.

Since the publication of the LB-2 and CAFASP2 reports, new servers have been developed and evaluated in subsequent LB rounds. To obtain an updated snapshot of the predicting capabilities of current servers, we report the main results from the recently completed LB-4.

LiveBench-4

LB-4 was run between November 2001 and April 2002. Seventeen servers were evaluated using 108 newly released structures as targets. For detailed evaluation information, including separation between 'easy' and 'hard' targets, and the four different evaluation methods used, see the LB-4 web pages. Here we report the LB-4 results using a simplified approach. Table I lists the performance of the LB-4 servers, using a rough ranking based on overall sensitivity. Sensitivity is defined here as the percentage of targets for which a (partially) correct prediction is obtained, as assessed by the MaxSub evaluation method (Siew *et al.*, 2000). We deliberately group similarly performing servers in one single range to emphasize the fact that the differences among servers may not be very significant and can be affected by the particular way of evaluation. We should also note that some servers (e.g. SFAM) are aimed at a particular type of prediction problems and thus may not be fairly ranked by an overall number.

Table I shows that four of the five top-performing servers identified in previous experiments still rank at the top (3DPS, INBG, FFAS and MGTH). However, four new servers now accompany them. FUG2 and FUG3 are variants of FUGU developed by the same group. ORFs is a profile-to-profile comparison server developed at bioinfo.pl by one of the authors of this paper. SHGU (3D-SHOTGUN) is an enhancement over the INBG method, developed by the other author of this paper (Fischer, 2003). PDBb corresponds to a local implementation of PSI-BLAST. It is clear that all FR servers outperform PDBb.

The sensitivity of the top ranking servers is above 50% (Table I). Apparently, there has been a slight improvement in the sensitivities of the 'old' LB-2 servers. However, this improvement may not necessarily mean that these servers have improved. It may merely be due to the differences in the test sets (LB-4 may have included more 'easier' targets), to the growth of the sequence and structural databases or both.

In a special category, LB-4 also evaluated the performance of three meta-predictors: CNS2 (version 2 of pcons, described above), 3DS3 and 3DS5 (two meta-predictors using the 3D-SHOTGUN algorithm) (Fischer, 2003). We distinguish meta-predictors from individual servers by the type of input required: a meta-predictor cannot run independently, explicitly requiring as input the predictions of at least one other participating server. The meta-predictors' sensitivity ranges between 56 and 60%, confirming that the use of information from more than one server can result in increased performance. It seems that the performance of these meta-predictors do effectively represent an improvement.

Table II lists the overall specificities of the LB-4 servers. The specificity is computed as a percentage by dividing the total specificity calculated in the LB-4 web page by the total number of targets (108), times 100. In LB-4, the total specificity

Table II. LiveBench-4 specificity results

Specificity range (%)	Names of servers					
41–49	SHGU 21.9	FFAS 6.8	ORFs 6.0	INBG 21.8	FUG2 4.8	FUG3 5.2
34–36	SFAM 10^{-3}	3DPS 0.1	FUGU 5.2	ORFb 35.9	ST99 18.5	
25–32	PDBb 0.05	MGTH 0.6	GETH 0.6			

The specificity is computed as the 'total' specificity computed in LB-4 divided by the total number of targets (108). The numbers following the servers' names are rough estimates of their '5% confidence scores'.

is a rough estimate of the number of correct predictions a server produces while producing up to five incorrect ones. Table II also lists for each server the score of its fifth incorrect prediction. For example, the FUG2 server had a specificity of 44% (total specificity of 47.6, divided by 108 targets). Its 5% confidence threshold is 4.8. This means that FUG2 approximately produced 48 correct predictions plus five incorrect ones with scores higher than 4.8.

Six of the individual servers that ranked highest in sensitivity also rank at the top in specificity (SHGU, FFAS, ORFs, INBG, FUG2 and FUG3). Surprisingly, in LB-2 MGTH ranked at the top in specificity, whereas in LB-4 it performs significantly worse. This may be due to a change in MGTH's scoring system introduced during LB-4. As expected, the servers' specificities are lower than their sensitivities. For example, while FUG2's sensitivity was 50% (correctly predicting 54 out of the 108 targets), its specificity was only 44%. That is, some of FUG2's correct predictions had scores lower than its 5% confidence threshold.

The meta-predictors' specificities are in the range 50–56%, again higher than the individual servers, most probably representing a significant improvement. The 5% confidence scores of 3DS3, CNS2 and 3DS5 are 24.7, 1.2 and 4.2, respectively. The 5% confidence scores reported in Table II are of great value to the users of the servers, as they give an indication of when a particular prediction may be reliable. Depending on the way in which each server scores its predictions, a more confident prediction can mean a larger or smaller score. Probably the most valuable outcome of LB, rather than attempting to provide an exact ranking of the servers, may be the availability of these confidence thresholds. The LB-4 web pages include detailed information listing the scores of the first to tenth false positives of each server.

As specified above, one of the advantages of LB is that it uses a relatively large test set (108 targets) compared with that used in CAFASP (at most one or two dozen FR targets). Thus, the relative ranking of the servers and the specificity analysis provided by LB are more robust than those produced at CAFASP. Despite the smaller test set used in CAFASP, both experiments identify almost the same group of top-performing servers.

The 2002 Structure Prediction Games

The results of the ongoing LB-6 and CAFASP-3 experiments will be particularly interesting because of the presence of new servers that have not been evaluated in LB-4. Another interesting addition in 2002 is the introduction of two new evaluation experiments, PDB-CAFASP and MR-CAFASP. The goal of PDB-CAFASP, like CAFASP3 and LB, is to evaluate the performance of fully automatic 3D protein structure prediction servers. The difference in PDB-CAFASP is the set of targets used. CAFASP uses CASP targets, LB uses newly released PDB entries, whereas PDB-CAFASP uses as targets pre-release PDB entries. Pre-release PDB entries are entries soon to be released, whose structures are not yet published, but whose sequences are known. Thus, PDB-CAFASP is, like CAFASP, a blind prediction experiment, where the predictions are made before the experimental structures are available.

MR-CAFASP is an experiment aimed at evaluating the potential of predicted models to be of aid during the experimental structure determination process. The focus will be on targets with no close homologue of known structure, where molecular replacement (MR) techniques cannot easily be applied, either because a parent of known structure is not easily identified or because the sequence–structure alignment is not reliable.

During previous LB experiments, it was discovered that in at least two cases, highly confident predicted models significantly differed from the experimental structure. In both cases, the experimental structure was subsequently removed from the PDB and, in one case, the replacement entry contained corrected models very similar to the original *in silico* prediction (Bujnicki *et al.*, 2002a,b). This lead to the consideration of whether *in silico* models may be of help during the structure determination process – a very important issue for structural genomics. Jones has also recently suggested that distant homology fold-recognition models may be used as molecular replacement phasing models (Jones, 2001).

The questions that MR-CAFASP aims to address are as follows. Would a predicted model be of help to fit the chain better into a low-resolution, hard to interpret, electron-density map? Can a predicted model help detect shifts and errors in the initial tracing of an electron-density-map? Can a predicted model be used as a phasing model? How can NMR benefit from an accurate predicted model? Because many predicted models may not be accurate enough, is it worthwhile for the experimentalist to spend some time verifying this?

In MR-CAFASP, highly confident fully automatic predictions will be selected from the targets of PDB-CAFASP, before the experimental structure is released. From these, a number of tests will be carried out *vis-à-vis* the experimental data, when the latter become publicly available.

For more information about PDB-CAFASP and MR-CAFASP, see the main CAFASP3 web page.

Developers and the users of structure prediction programs will again be watching the 2002 protein structure prediction Games. The community will learn whether new servers, not evaluated in LB-4, will demonstrate further progress or whether the main findings of LB-4 will simply be confirmed. More than ever, it will be interesting to see how useful the automated predictions will be for human CASP5 predictors and, more important, to see the performance differences between the current best automated servers and the best human predictors.

Disclaimer

It is important to state that the authors of this paper are at the same time the organizers of CAFASP and LiveBench and the developers of participating servers. Thus, even if the evaluation is carried out by fully automated procedures, it cannot be considered as independent. However, the generally consistent ranking obtained at CASP provides a valuable and independent control.

Note added after submission

The Asilomar CASP/CAFASP meeting was held shortly before this paper was accepted for publication. The results of the meeting largely confirmed the findings of LiveBench: (i) four of the top-performing servers at CAFASP-3 were also at the top in LB-4; (ii) the performance differences of many servers are very slight and an exact ranking is not very meaningful; and (iii) meta-predictors perform significantly better than individual servers. Possibly one of the most important findings was that owing to the improved performance of the meta-predictors, the performance difference between the best human predictors and the best servers is narrowing. A full report of the 2002 Asilomar meeting will be published in a special issue of *Proteins*. For updated information see the CASP5 site at http://PredictionCenter.llnl.gov/casp5, the CAFASP3 site at http://www.cs.bgu.ac.il~dfischer/CAFASP3 and the Live Bench6 site at http://bioinfo.pl/LiveBench.

References

Abbott,A. (2001) *Nature*, **409**, 4.

Baker,D. and Sali,A. (2001) *Science*, **294**, 93–95.

Bujnicki,J.M., Elofsson,A., Fischer D. and Rychlewski,L. (2001) *Proteins*, Suppl. 5, 184–191; see also http://bioinfo.pl/LiveBench.

Bujnicki,J.M., Feder,M., Rychlewski,L. and Fischer,D. (2002a) *FEBS Lett.*, **525**, 174–175.

Bujnicki,J.M., Rychlewski,L. and Fischer,D. (2002b) *Bioinformatics*, **18**, 1391–1395.

CASP4 (2001) *Proteins*, Suppl. 5; see also http://Prediction center.lnl.gov.

Cristobal,S., Zemla,A., Fischer,D., Rychlewski,L. and Elofsson,A. (2001) *Bioinformatics*, **2**, 5.

Eyrich,V.A., Marti-Renom,M.A., Madhusudhan,M.S., Fiser,A., Pazos,F., Valencia,A., Sali,A. and Rost,B. (2001) *Bioinformatics*, **17**, 1242–1243.

Fischer,D. (2000) In Altman,R.B., Dunker,A.K., Hunter,L., Lauderdale,K. and Klein,T.E. (eds), *Proceedings of the Pacific Symposium on Biocomputing*. World Scientific, Hawaii, pp. 119–130.

Fischer,D. (2003) *Proteins*, in press.

Fischer,D. and Eisenberg,D. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 11929–11934.

Fischer,D. and Eisenberg,D. (1999) *Curr. Opin. Struct. Biol.*, **9**, 208–211.

Fischer,D., Baker,D. and Moult,J. (2001a) *Nature*, **409**, 558.

Fischer,D., Rychlewski,L., Elofsson,A., Pazos,F., Valencia,A., Rost,B., Ortiz,A.R. and Dunbrack,R.L.J. (2001b) *Proteins*, Suppl. 5, 171–183; see also http://www.cs.bgu.ac.il/~dfischer/CAFASP2.

Jones,D. (1999) *J. Mol. Biol.*, **287**, 797–815.

Jones,D. (2001) *Acta Crystallogr.*, **D57**, 1428–1434.

Karplus,K., Barrett,C. and Hughey,R. (1998) *Bioinformatics*, **14**, 846–856.

Kelley,L.A., MacCallum,R.M. and Sternberg,M.J.E. (1999) In Istrail,S., Peuznor,P. and Waterman,M. (eds), *RECOMB 99, Proceedings of the Third Annual Conference on Computational Molecular Biology.* ACM, New York, pp. 218–225.

Lundstrom,J., Rychlewski,L., Bujnicki,J. and Elofsson,A. (2001) *Protein Sci.*, **10**, 2354–2362.

Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) *Protein Sci.*, **9**, 232–241.

Shi,J., Blundell,T. and Mizuguchi,K. (2001) *J. Mol. Biol.*, **310**, 243–257.

Siew,N. and Fischer,D. (2001) *IBM Syst. J.*, **40**, 410–425.

Siew,N., Elofsson,A., Rychlewski,L. and Fischer,D. (2000) *Bioinformatics*, **16**, 776–785.

Received September 30, 2002; revised January 2, 2003; accepted January 7, 2003