

COMMUNICATION

The 2000 Olympic Games of protein structure prediction; fully automated programs are being evaluated *vis-à-vis* human teams in the protein structure prediction experiment CAFASP2

Daniel Fischer^{1,2}, Arne Elofsson³ and Leszek Rychlewski⁴

¹Bioinformatics, Department of Computer Science, Ben Gurion University, Beer-Sheva 84015, Israel, ³Stockholm Bioinformatics Center, 106 91 Stockholm, Sweden and ⁴International Institute of Molecular and Cell Biology, Ks. Trojdena 4, 02-109 Warsaw, Poland

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

In this commentary, we describe two new protein structure prediction experiments being run in parallel with the CASP experiment, which together may be regarded as the 2000 Olympic Games of structure prediction. The first new experiment is CAFASP, the Critical Assessment of Fully Automated Structure Prediction. In CAFASP, the participants are fully automated programs or Internet servers, and here the automated results of the programs are evaluated, without any human intervention. The second new experiment, named LiveBench, follows the CAFASP ideology in that it is aimed towards the evaluation of automatic servers only, while it runs on a large set of prediction targets and in a continuous fashion. Researchers will be watching the 2000 protein structure prediction Olympic Games, to be held in December, in order to learn about the advances in the classical ‘human-plus-machine’ CASP category, the fully automated CAFASP category, and the comparison between the two.

Keywords: CAFASP/critical assessment of fully automated protein structure prediction methods/protein structure prediction

While the determination of the complete genome sequences of various organisms has already become routine, the experimental determination of the 3D structure of the proteins encoded in these genomes continues to be a very laborious process. Several hundred thousand protein sequences are currently available in the public databases, but the number of currently available 3D protein structures is just over 10 000. This difference in sizes has been referred to as the ‘sequence-to-structure gap’. Despite worldwide efforts aimed at speeding up protein structure experimental determination, it has become clear that the sequence-to-structure gap is not likely to disappear soon, and that in many cases, only protein structure prediction ‘*in-silico*’ may help bridge the gap. The goal is to feed a computer with the amino acid sequences of the proteins encoded in a genome, let it crunch some numbers and at the end of a fully automatic process, produce the correct 3D shapes of the proteins. Despite significant advances in the last few years, current protein structure prediction methods are far from achieving this goal. To assess the progress in the field, computational biologists have devised a number of prediction experiments that can be regarded as the Olympic Games of protein structure prediction.

CASP

Every 2 years, the protein structure prediction community gathers around its most important event: the CASP (Critical

Assessment of Structure Prediction) (Moult *et al.*, 1999) blind prediction experiment, devised by John Moult of the University of Maryland in 1994. In CASP, a few dozen proteins of known sequence but unknown structure are used as prediction targets. Contestants are asked to file their predictions before the real 3D structure of the protein is experimentally determined. The predictions are filed using various methods known in the field as homology modeling, fold recognition or threading and *ab initio*. Subsequently, when the 3D structure is released, an assessment of the accuracy of the predictions is carried out. This protocol ensures that no participant knows the correct answer while running his/her programs and, thus, the submitted responses effectively reflect the state of the art of blind prediction at the time of the contest. The value of the CASP experiment is enormous: it discourages over-enthusiastic claims of the predictors and informs researchers outside the prediction community, including biologists and commercial companies, about the capabilities, limitations and progress of current structure prediction.

The fourth CASP event is currently under way, and over 150 predicting groups world-wide are expected to participate. As in previous CASPs, CASP4 will culminate with a meeting in Asilomar in December 2000. This year, a number of new evaluation experiments will take place in parallel to CASP4, which together with CASP4 will create a winter Protein Structure Prediction Pentathlon. In what follows we describe two of these new events.

CAFASP2, a new experiment: fully automated structure prediction

Because in the CASP protocol human intervention is allowed when producing the predictions, one of its limitations is that it measures the performance of computer-aided structure prediction; that is, CASP measures the capabilities of human experts using prediction programs and not the capabilities of the programs themselves. However, assessing the performance of fully automatic methods is critical for biologists. When biologists aim to predict the structure of a protein, what they wish to know is which program performs best and not which group was able to produce the best predictions at CASP. With the advent of genome sequencing projects, including that of the human, the need for fully automated structure prediction has become evident. A few years ago, automated tools were either non-existent or highly inaccurate. But as protein structure prediction has evolved, and a number of automated tools have demonstrated that they are already able to produce valuable predictions in many cases, it is now important to test their capabilities alone. To address this, the Critical Assessment of Fully Automated Structure Prediction experiment (CAFASP) (Fischer *et al.*, 1999), was initiated in 1998 by Daniel Fischer of the Ben Gurion University in Israel. In CAFASP, the participants are programs or Internet servers and what is evaluated here are the automated results of the programs without any human intervention. CAFASP1 was a small

Table I. Protein structure prediction servers registered at CAFASP2

Server name	Type ^a	url (America)
PHD/PROF (Rost)	SS/SS	http://dodo.cpmc.columbia.edu/predictprotein
SSpro	SS	http://promoter.ics.uci.edu/BRNN-PRED
SDSC1	HM	http://c1.sdsc.edu/hm.html
FFAS	FR	http://bioinformatics.burnham-inst.org/FFAS
SAM-T99	FR/SS	http://www.cse.ucsc.edu/research/compbio/MMM-apps/T99-query.html
P-Map	FR	http://www.dnamining.com
loopp	FR	http://ser-loopp.tc.cornell.edu/loopp.html
123D+	FR	http://www.lmmb.ncifcrf.gov/~nicka/123D+.html
FORESST	FR	http://abs.cit.nih.gov/foresst2/
Isites	AI	http://honduras.bio.rpi.edu/~isites/ISL-rosetta.html
Dill-Ken	AI	http://www.dillgroup.ucsf.edu/~kdb
Cahill-Flehart	AI	http://kevin.phys.unm.edu/protein
Server name	Type ^a	url (Asia)
PSSP	SS	http://imtech.ernet.in/raghava/pssp
FAMS	HM	http://physchem.pharm.kitasato-u.ac.jp/FAMS
rpfold	FR	http://imtech.chd.nic.in/raghava/rpfold
Server name	Type ^a	url (Europe)
Jpred2	SS	http://jura.ebi.ac.uk:8888
Pred2ary	SS	http://www.cmpharm.ucsf.edu/~jmc/pred2ary/
PROF (King)	SS	http://www.aber.ac.uk/~phiwww/prof/index.html
Nanoworld	SS	http://ftp.decsy.ru/nanoworld/DATA/PROGRAMS/DNETOENT/dne-to-ent.htm
3D-JIGSAW	HM	http://www.bmm.icnet.uk/people/paulb/3dj
GenTHREADER/PsiPred	FR/SS	http://www.psiPred.net
3D-PSSM	FR	http://www.bmm.icnet.uk/servers/3dpssm
FUGUE	FR	http://www-cryst.bioc.cam.ac.uk/~fugue
ssPsi	FR	http://130.237.85.8/~arne/sspsi
threadwithseq	FR	http://montblanc.cnb.uam.es
CORNET	CP	http://prion.biocomp.unibo.it/cornet.html
PDG_contact_pred	CP	http://grados.cnb.uam.es/pazos
Server name	Type ^a	url (Middle East)
bioinbgu	FR	http://www.cs.bgu.ac.il/~bioinbgu/
Server name	Type ^a	url (Australia)
Sausage	FR	http://rcs.anu.edu.au/~arussell/TheSausageMachine.html

^aType of server: SS = secondary structure prediction; HM = homology modeling; FR = fold recognition; AI = *ab initio*; CP = contacts predictions.

experiment with only a handful of fold-recognition servers. The results of CAFASP1 demonstrated that although in most cases human intervention resulted in better predictions, it was clear that several programs can already independently produce reasonable predictions. CAFASP2 is one of the new categories to be included in the 2000 Olympic Games of protein structure prediction. CAFASP2 will run in parallel with CASP4 and will use the same predicting targets as those used by humans at CASP4. Over two dozen automatic servers from five continents have already registered for CAFASP2 and several other groups are furiously tuning their methods and building automated Internet servers in preparation for CAFASP2. CAFASP2 will cover all aspects and methods of automated protein structure prediction including the one considered to be the most difficult: *ab initio*. The first fully automated computer servers for *ab initio* prediction are three of the CAFASP2 participants. Most members of the prediction community, and in particular the non-expert protein structure predictors in the wider biology and genetics communities, are waiting to learn how much progress has been achieved in automated structure prediction. Protein structure prediction servers registered at CAFASP2 are listed in Table I.

Live Bench going large-scale

Another parallel event that will have its results available to the wide community by the end of this year is the LiveBench experiment, lead by Leszek Rychlewski in Poland. LiveBench overcomes one of the limitations of both CASP and CAFASP: the relatively small number of prediction targets. LiveBench follows the CAFASP ideology in that it is aimed towards the evaluation of automatic servers only, but runs in a continuous fashion. The assessment is carried out over a large number of prediction targets compiled from newly released protein structures that are immediately submitted via the Internet to the participating servers. LiveBench-1 is currently under way, with only a handful of fold-recognition servers. Preliminary results (Rychlewski *et al.*, 2000) show that the best servers are able to produce correct models for between one-third and half of all newly released structures that show no sequence similarity to other proteins of known structure. Another parallel, large-scale evaluation project lead by Burkhard Rost of Columbia University is aimed at the evaluation of automated homology modeling and secondary structure prediction methods. The main contribution of such large-scale evaluations is, like CAFASP, to inform biologists about the current performance

of available automated servers; the main difference is that this evaluation is carried out in a continuous fashion and using a larger number of prediction targets.

The problem of model evaluation

Throughout the last few years, it has become clear that evaluating the accuracy of the predicted 3D models *vis-à-vis* the real structures is a difficult problem, given the diversity in methods, knowledge and data used to produce each model and prediction. In CASP, different criteria have been used for the assessment, partly automatic, partly involving human expertise and knowledge. Each criterion focused on different aspects of a 3D model. Evaluating how good a predicted 3D model has turned out to be a controversial sub-field of research.

In CAFASP and LiveBench, a single, objective, fully automated, quantitative and reproducible evaluation method is used. To this end, numerical measures that can be added over all predictions have been developed so that an estimation of the overall performance of each prediction method can be obtained (Siew *et al.*, 2000). One of these measures, named MaxSub, assesses the quality of a predicted model by searching for the largest subset of C α atoms in the model that superimpose well over the real structure of the protein. From this subset, a normalized score that reflects the quality of the superimposition is produced. Although finding such a subset is a hard problem, we have shown that heuristics provide an efficient solution with excellent results. However, automatic evaluation of predicted models is likely to continue to be controversial and is part part of our ongoing research (A.Elofsson, Cristobal,S., Zemla,A., Fischer,D., Rychlewski,L. and Elofsson,A., (2000) submitted).

The availability of automated evaluation measures allows large-scale evaluation experiments, such as LiveBench, to take place. It also allows full automation to be achieved in CAFASP2 and in LiveBench, both in the way the models are produced and in the way they are evaluated.

Man versus machine?

One of the most anxiously awaited results of CASP4 and CAFASP2 is the comparative analysis of the performance of humans (CASP4) with that of the automatic programs (CAFASP2). One particularly attractive feature this year is that because the automated predictions from the servers are available long before the filing deadline for the human predictions, human predictors can make use of the automated results when preparing their predictions (but not vice versa). The differences between the automated and the human predictions is expected to be smallest for the secondary structure prediction and the homology modeling methods, because it is in these methods that human predictors make the largest use of automatic components.

The performance comparison of humans versus machines will allow for the first time the amount of human intervention required in current interactive predictions to be objectively quantified. Understanding and analyzing the aspects of human expertise that lead to a better human performance will allow their future incorporation into automated programs; this is and will continue to be one of the major challenges for developers. Comparing human and machine performance is beginning to raise interest similar to that observed for the man versus computer matches in chess. It took over 20 years of computer chess tournaments before a machine beat a grandmaster. As a sequel to the latter achievement, IBM has now joined the

protein structure prediction community with the development of their 'Blue Gene' project. Although machines will probably not outperform humans this year, we should not bet high against machines in CAFASP5.

Why automate structure prediction?

There are considerable benefits to be gained from studies of automated structure prediction, despite the criticisms that some researchers have raised. Similar doubts were raised when computer chess programs emerged. As in chess, the goal is not to replace humans, but to encourage further development of the automated tools, so they become more routine companions in the prediction tasks, ridding humans from as many tedious computations as possible, and allowing them to apply their intuition and expertise better. Another important goal of tournaments of automated tools is to allow biologists to choose the best performing tools for their particular prediction needs. If something is computable, programs should be written to compute it, and their performance should be thoroughly tested. The challenge is to gain a better understanding of what is being computed when a protein folds *in vivo*. Finally, improvements in automated structure prediction will allow us to distinguish more and more cases of accurate and reliable predictions. This will leave fewer cases for human intervention, a most important goal in the post-genomic era.

The CAFASP and LiveBench experiments are a contribution towards the long sought-after goal of being able to submit to a computer the complete genome sequence of an organism, and upon a number of calculations, obtaining the 3D structures of each of the encoded proteins. The result will be a major step towards our ability to understand the relationship between structure and function in biological systems, to prevent and cure disease and to control processes in living systems. Although this goal will not be fully achieved in CAFASP2, subsequent tests will serve as catalysts and measures of continuing success. The protein structure prediction community and the wider community of bioinformaticians and biologists using these tools will certainly be watching the 2000 protein structure prediction Olympic Games for the advances in the classical 'human plus machine' CASP category, for the new reports of the fully automated CAFASP category and for the comparison between the two.

For more information, see the CASP site at <http://Prediction-Center.llnl.gov/casp4>, the CAFASP site at <http://www.cs.bgu.ac.il/~dfischer/cafasp2> and the LiveBench site at <http://bioinfo.pl/LiveBench>.

References

- Bujnicki,J.M., Elofsson,A., Fischer,D. and Rychlewski,L. (2000) submitted.
- Fischer,D. *et al.* 1999; CAFASP1, *Proteins*, Special Issue. **Suppl. 3**, 209–217.
See <http://www.cs.bgu.ac.il/dfischer/cafasp1/cafasp1.html>.
- Moult,J., Hubbard,T., Fidelis,K. and Pedersen,J.T. (1999) *Proteins*, Suppl., 2–6.
- Siew,N., Elofsson,A., Rychlewski,L. and Fischer,D. (2000) *Bioinformatics*, in press.

Received July 31, 2000; revised August 8, 2000; accepted August 10, 2000