PROTEINS: Structure, Function, and Genetics Suppl 5:171–183 (2001)
DOI 10.1002/prot.10036

CAFASP2: The Second Critical Assessment of Fully Automated Structure Prediction Methods

Daniel Fischer,[1]* Arne Elofsson,[2] Leszek Rychlewski,[3] Florencio Pazos,[4†] Alfonso Valencia,[4†] Burkhard Rost,[5] Angel R. Ortiz,[6] and Roland L. Dunbrack, Jr.[7]

[1]*Bioinformatics, Department of Computer Science, Ben Gurion University, Beer-Sheva, Israel*
[2]*Stockholmm Bioinformatics Center, Stockholm University, Stockholm, Sweden*
[3]*BioInfoBank Institute, Poznan, Poland*
[4]*Protein Design Group, CNB-CSIC, Cantoblanco, Madrid, Spain*
[5]*CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York*
[6]*Department of Physiology and Biophysics, Mount Sinai School of Medicine of the New York University, New York, New York*
[7]*Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, Pennsylvania*

ABSTRACT **The results of the second Critical Assessment of Fully Automated Structure Prediction (CAFASP2) are presented. The goals of CAFASP are to (i) assess the performance of fully automatic web servers for structure prediction, by using the same blind prediction targets as those used at CASP4, (ii) inform the community of users about the capabilities of the servers, (iii) allow human groups participating in CASP to use and analyze the results of the servers while preparing their nonautomated predictions for CASP, and (iv) compare the performance of the automated servers to that of the human-expert groups of CASP. More than 30 servers from around the world participated in CAFASP2, covering all categories of structure prediction. The category with the largest participation was fold recognition, where 24 CAFASP servers filed predictions along with 103 other CASP human groups. The CAFASP evaluation indicated that it is difficult to establish an exact ranking of the servers because the number of prediction targets was relatively small and the differences among many servers were also small. However, roughly a group of five "best" fold recognition servers could be identified. The CASP evaluation identified the same group of top servers albeit with a slightly different relative order. Both evaluations ranked a semiautomated method named CAFASP-CONSENSUS, that filed predictions using the CAFASP results of the servers, above any of the individual servers. Although the predictions of the CAFASP servers were available to human CASP predictors before the CASP submission deadline, the CASP assessment identified only 11 human groups that performed better than the best server. Furthermore, about one fourth of the top 30 performing groups corresponded to automated servers. At least half of the top 11 groups corresponded to human groups that also had a server in CAFASP or to human groups that used the CAFASP results to prepare their predictions. In particular, the CAFASP-CONSENSUS group was ranked 7. This shows that the automated predictions of the servers can be very helpful to human predictors. We conclude that as servers continue to improve, they will become increasingly important in any prediction process, especially when dealing with genome-scale prediction tasks. We expect that in the near future, the performance difference between humans and machines will continue to narrow and that fully automated structure prediction will become an effective companion and complement to experimental structural genomics. Proteins 2001;Suppl 5:171–183.** © 2002 Wiley-Liss, Inc.

Key words: fully automated structure prediction; fold recognition; critical assessment; CASP; LiveBench; CAFASP; automated structure prediction evaluation

INTRODUCTION

In this postgenomic era, structure prediction, as many fields in modern biology, is undergoing a radical change: it is being transformed from being an art mastered by only a few expert-artists to a computational research area being applied by many non-expert predictors. The need for automatic structure prediction has never been more evident, as researchers realize that in the foreseeable future not all the protein structures will be solved, despite the number of worldwide structural genomics initiatives.[1] Paradoxically, as the number of known structures increases, the number of sequences that biologists expect to model increases. The utility of structural genomics will be achieved only if automated, reliable tools succeed to model most of the proteins closely and distantly related to proteins of known structures. What non-expert biologists need is to be able to apply automatic tools for their prediction needs, and on a large, genomic scale. In addition, as we gain understanding in protein modeling, it has become clear that there is little use of the expertise of

†Grant sponsor: Spanish Ministry for Science and Technology.

*Correspondence to: Daniel Fischer, Bioinformatics, Department of Computer Science, Ben Gurion University, Beer-Sheva, Israel 84015. E-mail: dfischer@cs.bgu.ac.il.

Received 9 March 2001; Accepted 20 September 2001

human artist predictors, if it cannot be reproduced and automated. Assessing the performance of automated structure prediction is thus essential to learn what the capabilities and limitations of the methods alone are. In addition, learning how much better than programs human expert predictors are is important to identify further directions of improvements in the automated methods.

We present here the results of the second Critical Assessment of Fully Automated Structure Prediction (CAFASP2). CAFASP was created in 1998,[2] as a result of the realization of many participants of CASP3[3] that a significant amount of human intervention was involved in the prediction process. Consequently, it became clear that the CASP experiments[3] were assessing the performance of human teams using programs and that it was not possible to measure the capabilities of the programs alone. CAFASP1 was a small experiment carried out after CASP3, with only a handful of participating fold recognition servers. It was a useful experiment that helped pave the way toward CAFASP2.

In CAFASP2 the participants are fully automatic web servers covering various aspects of protein structure prediction, and the assessment is carried out over automatically produced models without the human expert intervention allowed (but not required) at CASP. CAFASP is a parallel experiment to CASP, run on the same prediction targets as those of CASP4. Thus, the CASP/CAFASP marriage provides the unique opportunity of being able to compare models produced by human experts with those produced by fully automatic tools. To avoid the possibility that a particular human group may perform better at CASP only because it had better access to available servers, all the automated predictions of the CAFASP servers were made publicly available long before the human predictions were filed to CASP. Thus, CAFASP allowed human predictors to analyze, use, and possibly improve the servers' results when filing their predictions. Their challenge was thus to produce more accurate models than those produced by the servers. Because of this, a direct comparison of the predictions filed by programs versus those filed by human groups cannot be completely fair, but it still enables us to provide an indication of how much human expert intervention may contribute to a better prediction.

A secondary goal of CAFASP2 was to achieve full automation also in the assessment process. This led to the development of evaluation methods that can be applied in large-scale experiments.[4] Thus, each model submitted to CAFASP underwent two independent evaluations: one carried out by the CAFASP automated methods and the other by the CASP human assessors. These independent evaluations also provide a unique opportunity to assess the capabilities of automated evaluation tools.

Last, but not least, CAFASP is also extremely valuable for the users of the servers; it provides an indication of the performance of the methods alone and not of the "human plus machine" performance assessed in CASP. This information may aid non-expert users in choosing which programs to use and in evaluating the reliability of the programs when applied to their specific prediction targets.

Our full automation goals are not meant to belittle or to cast any doubt on the importance of the specialized expertise in structure predictions nor in human assessment capabilities. This work does not attempt to show that automated prediction is better or more desirable than "human expert plus machine" predictions. We believe that a knowledgeable human will—for the foreseeable future—do better (when using his expertise and time to interpret the automated method's results) than the automated method's results alone. However, whatever is computable by humans, if valuable, should be computable by machines, so that it can be scalable and reproducible by others. The challenge for bioinformaticians is to bring the human expertise, when possible, into programs that can be used by the wide community of users. Thus, the parallel assessment of programs and human groups in the CASP/CAFASP experiments is likely to result in significant advances in the field.

MATERIALS AND METHODS

All the methodology, predictions, and evaluation results are available through CAFASP's web site at http://www.cs.bgu.ac.il/~dfischer/CAFASP2. We present a brief summary of the methodology applied in the following.

Automated Servers and Prediction Categories

More than 30 servers, covering the five prediction categories of structure prediction, registered at CAFASP2 (Table I). The categories with the largest number of participating servers were fold recognition and secondary structure prediction, with 19 and 8 registered servers, respectively. The other three categories, namely, contacts prediction, ab initio, and homology modeling, had two or three registered servers each. Brief descriptions of a selected number of servers are included in the corresponding sections below.

Targets

CAFASP2 ran in parallel with CASP4, using the same prediction targets. Targets were classified into two main categories: homology-modeling (HM; 15 targets) and fold recognition (FR; 26 targets). This classification was based on whether PSI-BLAST[5] found good matches to proteins of known structure. If on convergence PSI-BLAST found a hit to a PDB entry with a score better than 0.001, then the target was considered to be an HM target; otherwise, it was considered to be an FR target. All targets were used as queries for all servers.

Filing the Predictions

On release of a prediction target, the CAFASP meta-server (http://cafasp.bioinfo.pl) submitted the amino acid sequence as a query to each of the CAFASP participating servers. The servers' replies were compiled during the following 48 h, and these replies were stored at the meta-server's site. Servers that failed to provide results within the 48 h were allowed to submit "late" predictions, but these predictions were not considered valid, nor were they taken into account in the evaluation.

TABLE I. Protein Structure Prediction Servers Registered at CAFASP2[†]

Fold Recognition	
FFAS	http://bioinformatics.burnham-inst.org/FFAS
SAM-T99	http://www.cse.ucsc.edu/research/compbio/
P-Map	http://www.dnamining.com
loopp	http://ser-loopp.tc.cornell.edu/loopp.html
123D+	http://123D.BioInfo.PL/run123D+.html
rpfold	http://imtech.chd.nic.in/raghava/rpfold
M/GenTHREADER	http://www.psipred.net
3D-PSSM	http://www.bmm.icnet.uk/servers/3dpssm
FUGUE	http://www-cryst.bioc.cam.ac.uk/~fugue
ssPsi	http://130.237.85.8/~arne
threadwithseq	http://montblanc.cnb.uam.es
bioinbgu	http://www.cs.bgu.ac.il/~bioinbgu/
Sausage	http://rcs.anu.edu.au/~arussell/TheSausageMachine.html
PDB-Blast	Psi-Blast run at http://bioinformatics.ljcrf.edu/pdb_blast
Secondary Structure Prediction	
PHD/PROF (Rost)	http://dodo.cpmc.columbia.edu/predictprotein
SSpro	http://promoter.ics.uci.edu/BRNN-PRED
SAM-T99	http://www.cse.ucsc.edu/research/compbio/
PSSP	http://imtech.ernet.in/raghava/pssp
Jpred2	http://jura.ebi.ac.uk:8888
Pred2ary	http://www.cmpharm.ucsf.edu/~jmc/pred2ary/
PROF (King)	http://www.aber.ac.uk/~phiwww/prof
Nanoworld	N.A.
Psipred	http://www.psipred.net
Contacts Prediction	
CORNET	http://prion.biocomp.unibo.it/cornet.html
PDG_contact_pred	http://montblanc.cnb.uam.es:8081/pdg_contact_pred.html
Ab initio	
Isites	http://honduras.bio.rpi.edu/~isites/ISL_rosetta.html
Dill-Ken	http://www.dillgroup.ucsf.edu/~kdb
Homology Modeling	
SDSC1	http://c1.sdsc.edu/hm.html
FAMS	http://physchem.pharm.kitasato-u.ac.jp/FAMS_S
3D-JIGSAW	http://www.bmm.icnet.uk/people/paulb/3dj

[†]For further details see http://cafasp.bioinfo.pl/server/

The servers' predictions were made available at all times at the meta-server's site. Because the CASP deadline for submitting human group predictions was weeks after the release of each prediction target, human CASP participants could make extensive use of the servers' results. Notice that because of the fast growth rate of the sequence and structural databases, it was possible that between the time the servers' results were compiled and the time of the CASP submission deadlines, some of the targets became easier to predict. Despite this possibility, and to ensure full automation, we did not resubmit the targets to the servers by the CASP deadline, which resulted, in that for some cases, the human versus machine comparison was unfavorable for the servers.

For each target, up to five alternative models were allowed (corresponding to the top five ranks of the server's output). The score a server received for a target was the score for the top rank only, but all the submitted models were evaluated.

Automated Evaluation Methods

All evaluations in CAFASP were carried out by fully automated methods in each of the prediction categories. These methods, briefly described in the corresponding sections in Results, were announced *before* the experiment began, so all participants knew how they would be evaluated. All evaluation methods were made available via the Internet.

RESULTS

Because the evaluation of each CAFASP category focuses on different aspects of structure prediction, we present the CAFASP results separately for each of the five categories of servers. Each of the evaluations were carried out by the corresponding coordinators and the following are their independently written reports.

Fold Recognition and Alignment Accuracy

Evaluation method

The evaluation method used in CAFASP2 was MaxSub,[6] as stated by the published rules of CAFASP2 before the experiment began (see http://www.cs.bgu.ac.il/~dfischer/CAFASP2/evalr.html and http://www.cs.bgu.ac.il/~dfischer/CAFASP2/evalrfr.html). MaxSub identifies the maximum superimposable subset of C_α atoms of a model and an experimental structure and produces a single normalized

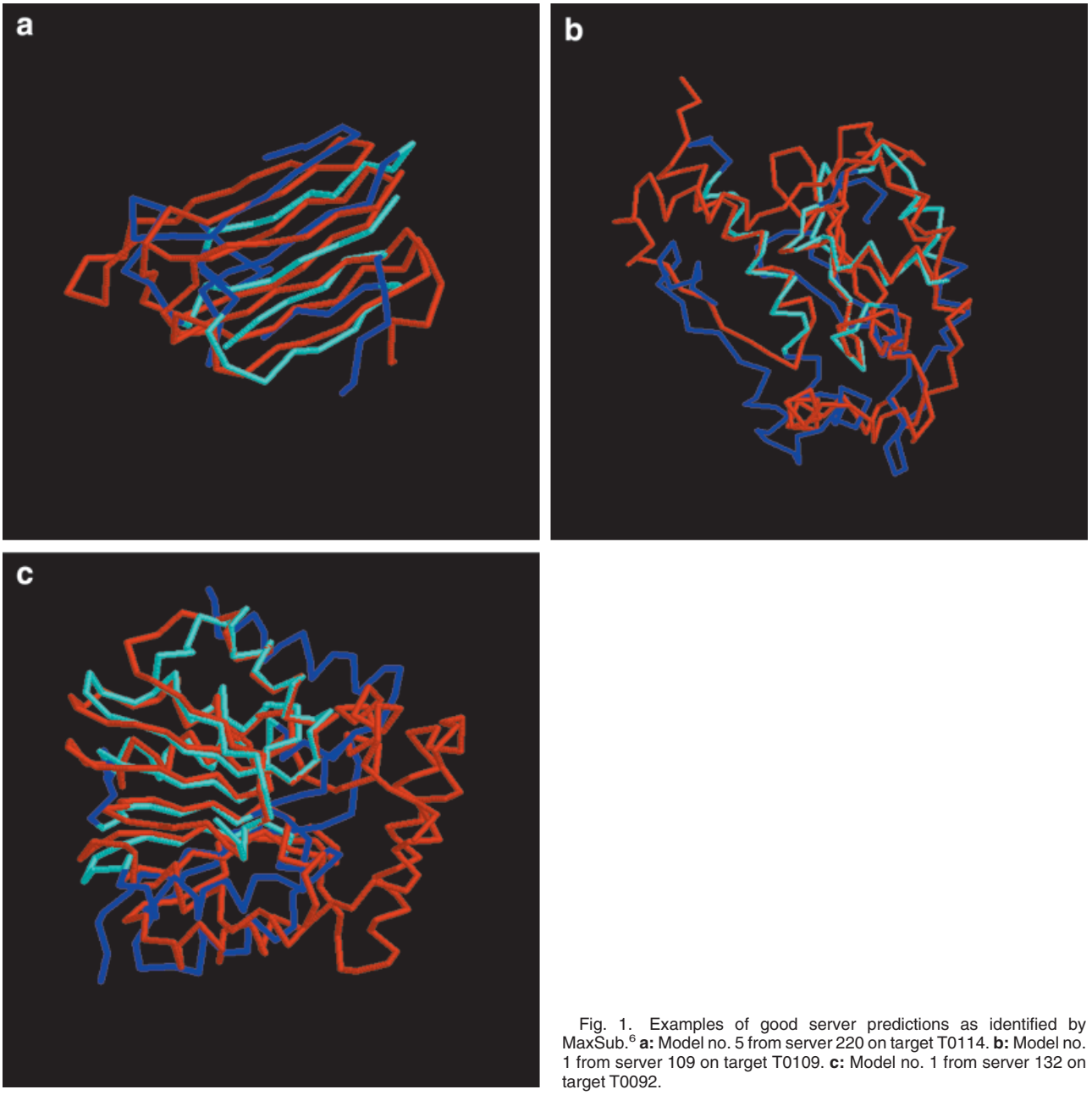

Fig. 1. Examples of good server predictions as identified by MaxSub.[6] **a:** Model no. 5 from server 220 on target T0114. **b:** Model no. 1 from server 109 on target T0109. **c:** Model no. 1 from server 132 on target T0092.

score that represents the quality of the model. MaxSub is a "sequence-dependent" assessment measure and produces scores in the range of 0.0–1.0 (Fig. 1), where 0.0 is an incorrect model, and 1.0 is a perfect model. A MaxSub score > zero was considered to be a correct prediction. MaxSub is available through the internet at http://www.cs.bgu.ac.il/~dfischer/MaxSub/MaxSub.html.

MaxSub has been extensively tested,[4,6] and a comparison between MaxSub's ranking and that of the assessments reported at CASP3[3] showed that good agreement was found for the more accurate models and for the better groups. The top five groups ranked by using the fully automated MaxSub were also the top five groups ranked at CASP3. Nevertheless, and as expected from any evaluation method, some differences were observed among the medium to poor models and groups, similar to the differences among the two CASP3 reports.[3] These differences may be due to the slight differences between the groups, to the different criteria used, to possible weaknesses in MaxSub, and to the subjectivity in the human assessment used in CASP. From this comparison and from the use of MaxSub in the LiveBench experiments[4] (see the LiveBench-2 report in this issue), we concluded that a measure such as MaxSub is suitable for the automatic evaluation of models.

As preannounced in the CAFASP2 rules published before the contest, "participation in CAFASP-2 implied acceptance of these procedures and rules." After the CAFASP2

TABLE II. Main CAFASP2 Evaluation Results

Target set	No. targets	Max. correct[a]	RANK First	Second	Third
Homology Modeling Targets	15	125	BIOINBGU (093 106 107[b]) 3D-PSMM MGenThreader	BIOINBGU (108) SAM-T99 GenThreader	FUGUE FFAS
Fold Recognition Targets	26	5	FFAS BIOINBGU (093 106 108)	FUGUE GenThreader	3D-PSMM MGenThreader

[a]The maximum number of correct predictions obtained by an individual server. A number of servers filed "late" predictions after the allowed 48-h period. These were not taken into account in the above evaluation. However, if these late predictions were considered, then a number of servers would rank close to the servers listed in the table. A case in point is that of the 123D+ server, which had a number of very good (but "late") predictions (see the CAFASP web site for details).
[b]The BIOINBGU server reports five different results, one for each of its components; the numbers in parenthesis give the CASP ID of the corresponding components.

rules were announced, other fully automated evaluation methods have been developed. These include the sequence-dependent and -independent measures called lgscore and lgscore2[4] and an experimental contact map overlap measure called *touch* (unpublished, but see http://cafasp.bioinfo.pl/touch/ for more details). For comparison and verification purposes, but not for official assessment purposes, these new additional evaluation methods were also applied. In addition, and only for comparison purposes, we also applied a CAFASP1-like evaluation and other variations of MaxSub by using different thresholds and normalizations, or considering the best of the five models submitted. All these additional evaluations (not shown) arrived at very similar results to those reported below. Because the main aim of this report is to assess the performance of prediction servers, and not the performance of evaluation methods, here we concentrate on the former and refer the interested reader to CAFASP's url at http://www.cs.bgu.ac.il/~dfischer/CAFASP2/ALevaluation/additional.html, where full details of the additional evaluations are listed. For a detailed comparison of many evaluation methods see ref. 37.

Only models number one for each target and server were considered in the official CAFASP-2 evaluation. For each category (HM and FR) the total score of a server was the sum of the MaxSub scores received for each target. To smooth the effect of a small number of targets and to account for the fact that the differences between a number of servers is only slight, we computed the ranks that each server achieves by considering the N different subsets of N-1 targets both for the 15 HM targets and for the 26 FR targets. For each server, the best rank achieved in any of the N subsets was registered and is reported. This resulted in having more than one server at a given rank.

All servers producing sequence-structure alignments or coordinates for at least the C_α atoms were evaluated here.

Evaluation results

Two aspects of the servers' performance were evaluated: sensitivity and specificity.

Sensitivity results. Table II shows the top performing servers for the HM and FR targets. Detailed tables and evaluation results considering all models per target are available in the corresponding tables from our main web page. Table II shows that five servers (FFAS, BIOINBGU and some of its different components, FUGUE, m- and GenThreader and 3D-PSSM) appear in the first three ranks in both sets of targets, whereas SAM-T99 appears at the top only in the HM targets. The top ranking servers had no difficulty in producing correct models for all 15 HM targets. However for the FR targets, the best of the servers succeeded to produce correct models for only five.

Within each set of targets, and after analyzing the servers' results, a further division into easy and hard targets was carried out (see CAFASP's url). This division was carried out qualitatively and rather arbitrarily, mainly based on the number of servers having correct predictions with confident scores. To obtain a clearer picture of the servers' capabilities, we evaluated the servers' performance on the easy (5) and hard (21) FR targets separately. The results of this evaluation show that most of the points the servers received came from the success within the easy targets and that the largest number of correct predictions by any individual server among the hard targets was only two. The overall scores that the servers received came from up to six targets. That is, the predictions of all other targets were all considered incorrect and contributed no points to the score. This small number of correct answers makes the exact ranking of the individual servers difficult because of the relatively large variations that can appear in such a small set.

In summary, it is clear that according to the MaxSub evaluation criteria, good models were predicted for all 15 HM targets and for the 5 easy FR targets. However, there was much lower success among the 21 hard FR targets (which included 4 new folds). The main evaluation above considered on-time models no. 1 only. Consequently, good predictions obtained by the servers at ranks >1 cannot be appreciated. Examples of such good predictions are model no. 2 from server 132 on target T0108, model no. 3 from server 389 on target T0110, model no. 5 from server 220 on target T0114, and model no. 2 from server 108 on target T0121, among others (see Fig. 1). These server predictions are among the best predictions for the corresponding targets (including those filed by human groups). This illustrates that often servers are able to produce good predictions at higher ranks; obviously, the question is how

can these higher ranked predictions automatically be identified (see below).

Human versus machine comparison. The human groups predictions filed at CASP were not made available. Consequently, we could not evaluate the CASP predictions with our automated methods. Thus, the performance comparison of servers and humans presented here is mainly based on the CASP assessors' evaluations. which also evaluated the CAFASP results along with the human predictions filed at CASP. When the human predictions become available from the CASP web site, interested readers can obtain the results of the automatic evaluation by using our web facilities.

In addition, this parallel evaluation of the CAFASP results also enables us to compare the CAFASP automated evaluation with that of the CASP assessor. However, because of the different evaluation methods used and because of the slightly different sets of targets considered, these two evaluations are not directly comparable. For example, the second domain of target T0096 was considered in CAFASP to be a HM target, whereas the CASP fold recognition assessor included it in his evaluation. The CAFASP evaluation deliberately excluded two targets (T0095 and T0124) that could not be evaluated properly with our automated tools. Finally, the domain partitioning of some of the multidomain targets was different (the exact domain definitions and the targets considered at CAFASP are listed in the CAFASP web site). Nevertheless, a rough, qualitative comparison of the rankings assigned by these two independent evaluations may still be valuable, if the above differences and limitations are taken into account.

Table III shows the ranks and scores that the CASP assessor assigned to the CAFASP servers, along with the top performing human groups of CASP. The performance analysis of the CASP groups is presented elsewhere by the CASP assessor (see the assessor report in this issue). The top CASP groups are shown here for comparison purposes only. The names of the top performing servers are shown; all other groups are listed with their CASP id number. The last column of Table III shows the assigned ranks from the CAFASP evaluation as shown in Table II.

Although the predictions of the CAFASP servers were available to human CASP predictors before the CASP submission deadline, only 11 human groups performed better than the highest ranking server, 3D-PSSM. Nevertheless, the score of the top human group (41.0) is significantly higher than that of 3D-PSSM (24.5), indicating that the *best* human predictor at CASP clearly did much better than the *best* server. We notice that at least half of the human groups that ranked at the top 20 to 30 ranks corresponded to human groups that also had a server in CAFASP or to human groups that used the CAFASP results to prepare their predictions. Some of the human groups with servers succeeded in filing better predictions than the servers, and others (within ranks > 12) did not. In particular, the group in rank 7 corresponded to a semiautomated method that filed consensus predictions by using the CAFASP results of the servers. Notice that the CAFASP-CONSENSUS group was assigned the rank of 0 to indicate that its performance was superior to the listed rank 1; however, we used the "0" to also indicate that it was not a fully automated CAFASP server participant.

TABLE III. CAFASP2 Automatic Server Evaluation Versus CASP4 Assessor's Ranking of Human Groups and Servers

CASP rank[a]	CASP score[a]	Group ID[b]	CAFASP rank
1	41.0	354-human	
2	37.0	384-human	
3	34.0	94-human (45)	
4	33.5	126-human (12)	
5	33.0	31-human (19)	
6	30.5	88-human	
7	27.0	359-CAFASP-CONSENSUS	0
…			
12	24.5	132-3D-PSSM	3
…			
19	21.0	395-FFAS	1
20	17.5	106-BIOINBGU-seqpprf	1
22	16.5	259-GENTHREADER	2
23	16.5	93-BIOINBGU-Consensus	1
27	15.5	260-MGENTHREADER	3
31	14.5	103-FUGUE	2
38	12.5	108-BIOINBGU-prfseq	1
45	11.5	111-SAM-T99	
47	11.0	105-server	
48	10.5	107-server	
…		servers and humans	
90	1.5	158-PSI-BLAST	21
…		servers and humans	
113		455-server	
127		279-human	

[a]CASP rank and score as originally computed by the CASP assessor (modified to include the correct FFAS predictions).
[b]Group ID is the internal CASP id number assigned to each group. The names of the top servers only follow the ID number; for the others, "human" or "server" is appended to indicate whether the ID corresponded to a CAFASP server or to a human CASP group. The CAFASP-CONSENSUS group was not fully automated. To highlight this and to indicate that its performance was above the best CAFASP servers, its CAFASP rank is given as a "0." The three top human groups that also had a server registered at CAFASP are indicated by listing in parenthesis the rank their server achieved.

Finally, Table III shows that many CAFASP servers are able to produce more correct predictions than the standard sequence comparison tool PSI-BLAST, which ranks at the bottom (although not at the last rank) of the table. However, the superiority of the servers cannot be established on the basis of only their superior sensitivity. The servers' specificities need also to be evaluated and compared (see below).

Comparison of the CAFASP and CASP evaluations. Both the automated CAFASP evaluation and that of the human CASP assessor identified the same group of top performing servers (the top eight servers identified by the CASP assessor were ranked within the first three places in the CAFASP evaluation), and both evaluations ranked the CAFASP-CONSENSUS group above all individual servers. However, it is clear that a number of differences exist in the exact ranking of the top servers. As noted above, these differences are due to the different

evaluation procedures applied, to the different domain and target definitions, to the small size of the test set (no server had more than five correct predictions), and to the small differences in scores between the servers. For example, ranks 20 and 27 are separated by only up to a difference of 2 score points; 2 points could be achieved by a moderately correct prediction for a single target. Thus, the exact ranking at each narrow score range may not be very significant, and, consequently, slight differences are to be expected from independent evaluations. Furthermore, this also strongly suggests that rather than assigning gold, silver, or bronze medals to the servers, the group of top servers should be judged to have an approximately similar performance, with the various servers having different strengths for different targets. A similar conclusion has been reached by the large-scale LiveBench evaluation experiments,[4] confirming the suggestion that these "best" servers have a very similar performance.

Although the CAFASP and CASP rankings are not directly comparable, there is one interesting difference that may be worthwhile noticing: the difference of ranking of the server 132 (in CAFASP-2 it ranked third, but the CASP assessment ranked it first among the servers). By analyzing these and other differences, we discovered that some of them are attributable to the sequence dependency or independency of the methods used. Because of the shifts in the alignments, more score points are awarded by the more lenient sequence-independent evaluations of the CASP assessor and of lgscore-2 (see http://www.cs.bgu.ac.il/~dfischer/CAFASP2/ALevaluation/additional.html). A discussion on whether a sequence-dependent or a sequence-independent measure should be used is out of the scope of this article; we refer interested readers to the references listed above for more details.

Specificity results. We computed the specificity on FR targets for servers whose reliability score could be parsed from their output and which had at least one correct prediction before its third false positive. Table IV lists for each server the number of correct predictions with better scores than the scores of the first, second, and third false-positive predictions.

The magnitude of the scores of the first three false positives help the user of an automated method to determine the reliability of a prediction. For example, the table shows that based on the CAFASP data, a user should be careful when interpreting a FFAS top-hit with a score below 7. Table IV shows that FFAS, m-GenThreader, and GenThreader are the most specific servers, followed by BIOINBGU and 3D-PSSM. However, it is clear that the specificities of the servers were not high. This low level of specificity was also observed in CAFASP1 and is one of the main aspects that all servers need to improve.

Unfortunately, we could not compare the servers' specificities to that of humans because in CASP4, predictions were not filed with confidence scores.

Fold recognition evaluation discussion

We conclude that it is difficult to establish an exact ranking of the servers because the number of targets is relatively small, and the score differences between the servers is also small. Nevertheless, the different evaluations (including that of the CASP assessor) seem to coincide in identifying five servers performing better than the rest: FFAS, M/GenThreader, 3D-PSSM, BIOINBGU (and some of its components), and FUGUE. SAM-T99 appears to follow closely after the top servers, showing excellent performance in the HM targets. All servers in CAFASP2 performed poorly at the hard FR targets, but the relatively small number of targets does not enable us to draw more general conclusions. Servers for Homology Modeling were also evaluated here, and it seems that the FR servers produce more accurate sequence-structure alignments for the HM targets than the HM servers do. In general, most servers are significantly better than pdb-blast, even among the HM targets. However, no single approach is markedly superior to the others evaluated. These general observations coincide with the findings of LiveBench (see corresponding article in this issue).

In general, the specificities of the servers on the HM targets and on the five easy FR targets are good, but on the hard FR targets, the specificities need to be improved. This also was observered in the CAFASP1 experiment. Further comparisons between CAFASP1 and CAFASP2 are hard to make because the prediction targets appeared to be significantly harder in CAFASP2 and because in CAFASP1, only a limited evaluation was carried out. Now that a number of automated tools are in hand, we hope that for future experiments, objective and quantifiable comparisons to measure progress will be easier to produce.

TABLE IV. Specificities of the Servers[†]

Server	f1	f2	f3	HM
FFAS	3 7.2	4 6.81	4 6.33	14 —
mGenThreader	2 0.92	2 0.73	4 0.58	14 —
GenThreader	2 0.92	2 0.73	3 0.58	14 —
BIOINBGU (107)	2 6.6	2 5.3	2 5.3	14 —
3D-PSSM	2 0.285	2 0.298	2 0.403	14 —
BIOINBGU	1 22.9	2 20.5	3 19.0	14 —
BIOINBGU (108)	1 6.8	2 6.3	2 6.2	14 —
PDB-Blast	1 2e-07	2 0.015	2 0.033	1 e-103
BIOINBGU (106)	1 6.0	1 5.8	2 5.6	14 —
SAM-T99	1 2e-03	1 0.096	1 0.272	14 —
123D+	1 5.37	1 4.92	1 4.86	9 —

[†]Specificity computed on FR targets for servers with parsable reliability scores and with at least one correct prediction. Only model number 1 was considered per server and target. For each server, the number of correct predictions before the first, second, and third false positives are listed (columns f1, f2, and f3). The first number shown in each of the f1, f2, and f3 columns corresponds to the number of correct predictions with better scores (as reported by the server) than the score for the corresponding false-positive prediction. The second number corresponds to the score reported by the server. In this table, a prediction is defined as correct or wrong only (i) if three of the four evaluation methods (MaxSub, LGscore1, LGscore2, and Touch) agree (predictions where two methods say wrong and two methods say correct, are removed) and (ii) if the model returned by the server is longer than 25 residues (short predictions are removed). The column HM shows the total number of correct hits obtained before the first false positive on the HM targets (because of the above definition of "correct," only 14 of the 15 HM targets were considered in most cases). —, indicates there were no false positives to report.

In what follows we include commentaries from a group of server-developers. These were selected either because of their performance or because their methods appeared to be interesting or novel. The servers' developers were asked to also comment on how they used the servers' results in filing their human CASP predictions.

Selected fold recognition servers

3D-PSSM[7] uses sequence profiles from the structural superposition of remotely homologous structures together with standard sequence profiles, solvation terms, and secondary structure matching. Functional information, in the form of shared SWISS-PROT keywords, is used to postfilter predictions (SAWTED[8]). 3D-PSSM performed best overall according to the human expert CASP assessment. In the automated CAFASP assessment, 3D-PSSM performed first in the close homology section and came third in the remote homology section. However, the remote homology section of CAFASP had sparse data because of the lack of reliable automatic techniques for remote homology assessment. For the vast majority of the manual predictions in CASP filed by the Sternberg group, we used one of the top 10 matches from the 3D-PSSM server. A regularly updated fold library, textual information, and reference to other servers' results contributed to our CASP success (see our CASP report in this issue).

BIOINBGU[9] is a consensus method computed from five components, each exploiting evolutionary information in different ways, all based on sequence-to-structure compatibility using sequence-derived properties.[10] Fischer's CASP predictions were mainly based on BIOINBGU's predictions, resulting in only a slightly higher rank. The time difference between the CAFASP and CASP deadlines resulted in one better prediction. The server's results were considered correct only if high scores were obtained by a number of components, if similar results were obtained by other servers, or by confirmation from queries of homologues. This occurred for a handful of the FR targets only. For the others, the server was only marginally useful, requiring human intervention. This resulted in one correct and two incorrect decisions. Improvements in the way the consensus is computed are likely to result in a better performance.

GenThreader[11] and mGenThreader use position-specific scoring matrices (PSSMs) calculated by using PSI-BLAST[5] and a neural network based scoring system. The GenThreader algorithm has a bias toward superfamily level matches, and, as expected, the method clearly performs well on distant homology targets and poorly on hard fold recognition targets (analogous fold similarities). However, the method has a relatively high degree of selectivity. Jones's group CASP predictions were essentially based on a consensus of methods (including GenThreader and mGenThreader) and clues derived from biological function. The most pressing issue for improvement is to address the fact that GenThreader is only able to identify superfamily-level similarities.

FFAS,[12] fold and function assignment system, is a profile-profile alignment algorithm, developed to maximize the amount of information that can be extracted from a multiple alignment of a protein family *without* using any structural information. FFAS uses a two-dimensional weighting scheme that calculates the relative contribution of a specific sequence to the profile based on a matrix of similarities between all members of the family. It also uses a normalized scoring function that avoids overdependence on most conserved regions in the sequence and uses two directional scoring that explores a nonsymmetric nature of the profile-profile score.

New fold recognition servers

FUGUE[13] takes a (filtered) PSI-BLAST alignment and compares it against the structure-based alignments in the HOMSTRAD database.[14] The major differences from other methods include the use of well-parameterized environment-specific substitution tables and structure-dependent gap penalties. FUGUE currently does not use predicted secondary structures (which is an obvious element for improvements). In the CASP predictions filed by Blundell's group, FUGUE and other tools were used to compile candidate hits. FUGUE's alignments were manually examined and modified and fed to other tools for model building and evaluation. This resulted in better manual predictions. For example, in one case (T0108), manual adjustments of the input PSI-BLAST alignment led to a correct prediction, which was missed by the server. Work is in progress to automate some elements of the manual operations, such as model validation and its feedback to alignment.

LOOPP[15] evaluates the sequence-to-structure fitness by using two layers of contacts instead of the common one layer. The multilayer picture incorporates multibody effects. The parameters of the scoring function were optimized by linear programming, recognizing exactly the native structures of the training set. Emphasis is made on structural fitness, to detect remote homologs. Therefore, it is expected that LOOPP may miss targets recognizable by Psi-BLAST. Nevertheless, LOOP produced a number of interesting results including two of the difficult targets (T0097 and T0102).

Are the servers as a group useful?

When observing the number of correct predictions of all the servers as a group, we find that the servers identified roughly double the number of correct targets than the best of the servers. This indicates that each server appears to have its strengths in different targets and that the combined predictions of all servers can lead to a much better performance. The CAFASP organizers wanted to test the idea of how useful the combined results of the servers can be for a non-expert predictor and how a fully automated consensus prediction would compare with the individual servers and human groups. To this end, we aimed at a fully automated procedure that uses all the servers' results. Unfortunately, by the time CAFASP took place, our methods were not fully automated, and we had to register this consensus method as a CASP participant, named CAFASP-CONSENSUS. We emphasize that its success must be

attributed to each CAFASP participant who allowed us to use his/her server results. The performance of this CASP participant illustrates the utility of the servers' results as a group. We include here a description of the method used in CASP and of its newly developed fully automated version Pcons,[38] already available as a server (see the LiveBench-2 report elsewhere in this issue and http://www.sbc.su.se/~arne/pcons/).

CAFASP-CONSENSUS. The process of both the automatic Pcons server and the manual consensus predictions filed at CASP4 can be described in three steps. The first step is to detect structural similarities between predictions. In the manual procedure we listed the SCOP folds for each prediction. In the automated version the structures of the produced models are compared with one another. The second step is to identify related predictions. In the manual predictions we simply counted the number of predictions corresponding to each SCOP fold. In the automated version, the number of other models that are similar to a particular model is counted. The third step is to select one model. In the automated consensus predictor this is done by a neural network that combines the assigned score of a model with the information about the number of other similar models. For the manual consensus predictions we used the same information as input to our human neural networks. The CAFASP-CONSENSUS computations used at CASP4 were published at http://www.cs.bgu.ac.il/~dfischer/CAFASP2/ summaries before the submission deadlines of the targets. Thus, CASP predictors were able to use this information in filing their own predictions to CASP. In CASP4 we detected three scenarios. The first scenario consisted of trivial predictions, where most methods predicted the same fold. These predictions all turned out to be correct, including the "easy" FR targets. The second scenario was when no server produced a significant hit, but a particular SCOP fold was selected more frequently than all others. In the third scenario, no fold was significantly more frequent than the others. These predictions all turned out to be wrong.

Secondary Structure Prediction

Statistics about the results

Eight secondary structure prediction servers submitted predictions to CAFASP2 (see Table I). In total, we could analyze predictions for 40 proteins. Twenty-nine of the 40 proteins had no significant sequence similarity to any known structure (<25% pairwise identical residues in >100 residues aligned). For 2 of the 40 proteins an iterated PSI-BLAST search detected a clear similarity to a known structure; for 9 of the 40 the structural homologue was found by pairwise database searches. Most servers submitted results for most CASP targets; the highest number of submissions came from PROF_king, SAM-T99sec and SSpro, and the lowest number came from PROF-PHD. All results are available at: http://cubic.bioc.columbia.edu/mirror/cafasp_sec/. The particular details of the evaluation are explained in detail at http://cubic.bioc.columbia.edu/mirror/cafasp_sec/evaluation.html.

Results

Methods appeared more accurate than those at CASP2. All secondary structure prediction methods that were submitted to CAFASP2 submitted predictions for only 12 sequence-unique proteins. This set was too small to draw other conclusions than the following. (i) As expected, methods not using alignment information (PSSP) were less accurate than methods using evolutionary information. (ii) All methods using alignment information surmounted the accuracy we reached at CASP2; the best ones now appear to reach levels of 76% three-state per residue accuracy (percentage of residues correctly predicted in either of the states helix, strand, or other). (iii) The I-sites server was the only method evaluated that was not optimized to predict secondary structure; nevertheless, it predicted secondary structure rather accurately. All alignment-based methods were slightly worse when measuring the overlap of segments (SOV 9) and had similar levels of accuracy for helix and strand. Did methods perform better on proteins with homologues of known structure? The answer appeared affirmative for proteins with close sequence similarity to known structures and negative for proteins with only PSI-BLAST similarities to known structures. However, these statements were based on nine (close relation) and two (PSI-BLAST relation) proteins. Given these small sets, the differences in performance between proteins with structural homologues and proteins without were not significant.

How can we obtain significant comparisons between methods? The CASP4 assessors also evaluated secondary structure predictions, and to the best of our knowledge, their conclusions were similar to the ones obtained by us. Were these conclusions invalid due to the small data sets? The problem with small data sets is not that we cannot draw any conclusion. Rather, the task is to separate between significant and nonsignificant statements. One important issue at the CASP meetings is the ranking of methods. Neither the human expert CASP evaluation nor our automatic evaluation in CAFASP could nor did rank the methods in detail. However, users may assume that methods do in fact differ. To better rank the different methods, we need to evaluate all methods on equal data sets of significant size. Such an evaluation is described in this issue (see the EVA report in this issue).

Homology Modeling

Only two comparative modeling servers were entered in CAFASP2: FAMS and 3D-Jigsaw. These two servers produced complete models of targets, including insertions/deletions and side-chain coordinates. Servers that did not produce side-chain coordinates were not considered in this category.

FAMS ("Full Automated Modeling Server") developed by Ogata and Umeyama[16,17] begins with a Smith-Waterman alignment of the target sequence to structurally aligned template proteins with a substitution matrix derived from structure alignments.[18] C_α coordinates are obtained from the template proteins by a process maximizing "local space homology," or sequence similarity of residues within a

sphere of 12 Å radius to each segment to be modeled. The backbone model is then completed by borrowing from the template structures and simulated annealing. Conserved side-chains are kept in their crystallographic conformations, and the backbone model is adjusted to fit the side-chains of these residues. Other side-chains are then constructed by a procedure based on principal component analysis of local structural environments around side-chains in the PDB. Alternating cycles of Monte Carlo optimization of the backbone and side-chains are then used to optimize the predicted structure according to an objective potential function.

The 3D-Jigsaw server is an automated version of a homology modeling procedure of Bates and Sternberg that was successful at CASP3.[19] The backbone model is constructed from as many as five parent structures, whereas loops are constructed from fragment database searches and a mean-field algorithm for selecting the best conformations. Side-chains are constructed with the aid of a secondary structure-dependent rotamer library.[20]

Fold Assignments

To produce accurate homology models obviously requires correct fold assignment. In nine cases, the FAMS web server based its model on incorrect fold assignments, indicating a need for assessment of actual homology before proceeding with model building. By contrast, 3D-Jigsaw's fold assignments were uniformly correct, although it did not attempt targets with only very distant relatives in the PDB.

Backbone Models

Because alignments with parent structures were not provided by the servers, it was impossible to tell what loops contained insertions or deletions from the parent structures and therefore, were constructed ab initio. Therefore, the backbone models of these servers were assessed by comparing the backbone dihedrals ϕ and ψ with those in the experimentally determined structures for two sets of residues: those aligned by MaxSub within 3.5 Å of the experimental structure after structure alignment; and all residues in the model. These results are presented in Table I in our web site at http://www.fccc.edu/research/labs/dunbrack/cafasp-results.html. Backbone conformation for a single residue is considered "correct" if the value of D is < 60, where D is the root-mean-square deviation (mod 360) of the ϕ and ψ dihedrals for the residue from the experimental structure. The MaxSub results indicate that FAMS aligned correctly larger portions of each target sequence to template structures. The number of residues within 3.5 Å in the MaxSub alignments for several targets is much higher in the FAMS predictions than in the 3D-Jigsaw predictions. Of the correctly aligned residues, the rate of correct prediction of conformation does not differ substantially between the FAMS and 3D-Jigsaw predictions. This is to be expected because "correctly aligned" is correlated with reasonably correct backbone conformation. When considering all residues in the predicted structure, FAMS performed better than 3D-Jigsaw for most targets. Because FAMS structures contained more residues in total, the portions of the target protein (not just of the predicted portion) were also higher.

Side-chain conformations

Side-chain conformations were also compared with the experimental structures in two groups: those within 3.5 Å in the MaxSub structure alignment and all residues. The results are shown in Table II in our web site. Side-chain conformations were considered "correct" if $\chi 1$ was within 40° of the experimental structure values. The FAMS server produced somewhat better side-chain conformations than 3D-Jigsaw, although this may be due to the fact that larger portions of the structures were modeled correctly on the basis of the available parents. Most of the template-target sequence identities were well below 50%, and so the side-chain accuracy results are not particularly surprising.[21] It should also be noted that a backbone-independent rotamer prediction of $-60°$ for all side-chain types except Val (180°), Ser (+60°), Thr (+60°), and Pro ($-30°$) would achieve an accuracy rate of approximately 56%.

It is clear that the accuracy of homology models depends heavily on the initial alignment. Thus, we suggest that to obtain a detailed assessment of the side-chain and loop predictions, standard alignments be provided in future experiments, so that the capabilities of the methods in this aspect be better assessed.

Contacts Prediction

A detailed description of the evaluation for contacts prediction can be found at http://montblanc.cnb.uam.es/cnb_pred/abcp_eval.html. The evaluation has been implemented as a server that accepts as input a list of potentially contacting residues or a PDB file plus the associated confidence values for each pair of residues.

The results of the evaluation of contact prediction servers can be found at http://www.pdg.cnb.uam.es/cnb_pred/cafasp2_cp_eval/. Twenty-two predictions of CORNET server[16] and five of PDGCON (unpublished) were evaluated. To give an idea of the volume of data for the L/2 class, 2087 pairs of contacting residues predicted by CORNET were evaluated.

CORNET is based on previous work in training neural networks with family sequence profiles, under a carefully designed schema of sequence separation and protein size classes.[22] The new version incorporates information from correlated mutations,[23] sequence conservation (as in Ouzounis et al.[24]), and predicted secondary structure. PDGCON implements the original methods of correlated mutations[25] as in Pazos et al.[23] Even when it is recognized that contact prediction with correlated mutations can be improved by different procedures,[26] they are unfortunately too slow for their use in the server.

Despite the relatively small number of predictions evaluated in this first round, it is quite clear that the contact predictions based on a Neural Network (CORNET) are almost two-fold better ("Acc" value at 0.1xL of 0.22) than the simple correlated mutations implemented in the PDG-

CON server, but this level of accuracy is still relatively low, leaving significant room for future improvements. All the evaluation details are available at the url listed above.

Ab initio

Here we describe the evaluation of ab initio servers in CAFASP2. To better describe the type of methods used in this category, the ab initio category was renamed as "New Fold Methods."

Evaluation method

Model evaluation was based on a new method for structural alignment and comparison, which enables us to compare an experimental protein structure with an arbitrary low-resolution three-dimensional protein model (Ortiz and Olmea, submitted). The method, called MAMMOTH (http://transport.physbio.mssm.edu/services/mammoth), provides a structural similarity score between either two proteins or two different conformations of the same protein, derived from the likelihood of obtaining a correct fold prediction by chance. We briefly describe the four steps of the MAMMOTH algorithm.

1. From the C_α trace, compute the unit-vector U-RMS between all pairs of heptapeptides of both model and experimental structure.[27] This is a measure sensitive to the local structure.
2. Use the matrix derived in step 1 to find an alignment of local structures that maximizes the local similarity of both the model and the experimental structure. For that, use a global alignment method with zero end gaps.[28]
3. Find the maximum subset of similar local structures that have their corresponding C_α close in Cartesian space. Close is considered here as a distance $\leq$4.0 Å. The method to find this subset is a small variant of the MaxSub algorithm.[6]
4. Obtain the probability of obtaining the given proportion of aligned residues (with respect to the shortest protein model) by chance (P value). The P value estimation is based on extreme-value fitting.[29]

Results

Two different servers took part in CAFASP2: I-sites server (IS) and the Dill-Ken (DK) server.

I-sites is the first public ab initio protein structure prediction server. I-sites uses a very similar method to the one used for CASP3 by the Bystroff/Baker group, a combination of Bystroff's I-sites Library[30] and Baker's Rosetta algorithm,[31] scaled down in the public version. I-sites ranked at the top in the CASP3 ab initio category. I-sites did not use the structures of known homologs. Instead, it generates a sequence profile[5] to search for matches to the short sequence-structure motifs in the I-sites library. Where there are strong matches, the local motif structure is fixed by restraining the backbone torsion angles. Where there are multiple weak matches, those motifs become the moveset for a Rosetta conformational search.

ELAN-PROT (ELAstic Net prediction of PROTein structures) ab initio server uses the elastic net method to approximate the free-energy landscape of a protein by using deterministic annealing methods. ELAN-PROT uses a C_α-only model of residues connected by springs and includes the following non-local interactions: hydrophobic burial, a statistical interresidue potential, and restraints for secondary structure predictions from the PHD server.[32] The method is still relatively undeveloped and may be useful as a fast front-end to higher-resolution methods.

A total of 152 models were evaluated, belonging to 33 different targets. The IS server submitted 125 models, whereas the DK server provided 27 models. All models provided by the new folds methods servers were analyzed with MAMMOTH, and the results are summarized in the CAFASP2 web site. The most apparent finding is that neither of the two servers provided a single correct fold prediction. Score values >4.0 indicate substantial similarity between parts of the model and target, that is, a similarity of model and target that has very low probability to have been obtained by chance. Only 4 of the 152 models had a score larger than this threshold, all generated by the IS server, although 3 additional models from the IS server were close to the threshold and provided parts of the models correct. To obtain some additional perspective for these numbers, the best scoring models were compared by using MAMMOTH, with all models submitted to CAFASP2, either by fold recognition or ab initio folding. It was found that these models were among the best models in CAFASP2 (T0095TS216_4 ranked second overall, and T0097TS216_1 ranked fifth, whereas T0106TS216_1 ranked 3rd). Thus, it is encouraging to find that the best new fold methods predictions are also among the best models overall and correspond to some of the targets classified as "difficult" in CASP-4. We emphasize, however, that the detailed ranking values are not very significant when the structural similarities are so close to the random level. However, one of the goals of this experiment is not to determine the position of the ab initio prediction in the ranking of the threading models, but rather, to check whether the predictions are, with our score, of different qualitative nature between both approaches. This check is of interest for those cases where ab initio provides at least a correct partial structure, because it enables us to answer the question of "how far" the prediction is from that obtained on average with a threading server. A more detailed description of these data is available in our web page. In summary, although neither of the two serves provided a fully successful fold prediction, as assessed by the MAMMOTH scoring system, the IS server was able to predict for some difficult targets conformations of fragments with statistically significant scores.

DISCUSSION

CAFASP represents one of a number of current automation efforts aimed at meeting the challenges of the post-genomic era. CAFASP2 enables for the first time a direct comparison of the blind prediction capabilities of servers

with those of human teams. CAFASP2 showed that only a handful of human groups performed better than the servers, this despite the fact that the comparison was somewhat unfair toward servers (see Materials and Methods). In many cases, the human predictions required extensive human intervention involving a number of predictors in the group, applied additional biological knowledge or used the servers' results. The servers' performance in CAFASP2 is thus encouraging and signifies a bright future for automated structure prediction. However, it is important to notice that ranking high in CASP does not mean you are doing well; it only means you are doing as well as the best groups in the world. Even the best predictors at CASP produced accurate models for only a fraction of the harder targets. This means that there is still much room for improvement both in the human expertise and in the automated methods.

In addition, CAFASP proved to be a useful experiment in that (i) it allowed to measure how much better than servers humans are at predicting; (ii) it showed to the community of biologists and other predictors what the servers can and cannot do (a reality check); (iii) it was an open experiment where all was known and open to all: the participants and coordinators were known at all times, the evaluation rules and methods were defined before the experiment began, the evaluation methods were publicly available, all the predictions were available to all as soon as they were collected, and all the results were released as soon as they were obtained; (iv) it applied fully automatic, reproducible, quantitative, objective, and publicly available evaluation methods; (v) it enabled us to gain insights to improve future experiments such as CAFASP3 and the large-scale evaluation experiments LiveBench and EVA; and (vi) it helped to promote the development of automated tools that can be used by the wide community of users.

Assessing the performance of automated methods is of utmost importance to evaluate their applicability at genomic scales, in particular in the area of structural genomics.[1] Automated approaches for structure prediction are essential if the wealth of data in genomes is to be exploited (e.g., Refs. 1, 33, and 34). The CAFASP2 results indicate that progress in this direction has been made and that it seems that automated structure prediction has just begun to be ready to meet this challenge.

As a by-product of the CASP/CAFASP experiments, a new community wide effort, named The Most Wanted, has been launched.[35] This will be a joint effort of predictors that will attempt to tackle a number of important prediction targets with no experimental structure soon to be solved. Thus, rather than investing considerable time in competing on predictions of proteins that are soon to be solved anyway, many predictors will contribute their expertise on prediction targets that are relevant to the biological community. The automated servers for structure prediction will provide an invaluable infrastructure for this project, by producing an initial set of automated results that could be used by the expert predictors. Beyond this prediction effort, it is evident that automated structure prediction is effectively extending and reinforcing the impact of the current experimental structure genomics projects.[1]

Despite the success of the CAFASP2 experiment, it is important to notice that there is a long way to go before automatic structure prediction becomes a solved problem. We noted above that the success of the fold recognition servers among the harder FR targets is still modest and that the automated prediction of new folds is not yet accurate enough. Thus, there is still much room for significant improvements. It is evident that for the development of a new drug, extensive expert human intervention is still required, but as the automated methods are optimized, so is the efficiency of human experts' time. Another aspect of CAFASP that needs improvement is in the area of automatic evaluation. One of the main limitations of small-scale experiments, such as CASP/CAFASP, is the relatively small number of prediction targets. This requires that some caution is taken when interpreting the results of the experiment. To better assess the capabilities of servers, large-scale experiments, such as LiveBench and EVA, are required. These provide an important complement to the CASP/CAFASP experiments and together provide a better picture of the capabilities in the field.[36]

Finally, we hope that the success and prospects of CAFASP2 encourage many more developers of methods to fully automate their programs and to make them available through the Internet so that in future experiments we see a much larger number of server participants in each of the CAFASP prediction categories.

ACKNOWLEDGMENTS

We thank the CASP organizers and assessors for their interest and support; the many users of our servers and the many CASP participants and other supporters and advisors who encouraged us; the experimentalists that permitted their sequences to be used as benchmarks. Special thanks to Janusz M. Bujnicki for his contribution to the CAFASP-CONSENSUS predictions. Thanks also to Nir Esterman and Amir Mishali for their help in producing the summaries of the servers' results in the semiautomatic version of the CAFASP-CONSENSUS and to Naomi Siew for encouragement and discussions. The CNB-CSIC team acknowledges the computer facilities provided by the Supercomputing Center (Complutense University, Madrid).

NOTE ADDED IN PROOF

A number of new fold-recognition servers are currently being evaluated in LiveBench. Among others, there is a new version of pcons, pcons2, and a new meta-predictor named 3D-SHOTGUN (Fischer, in preparation). Visit the LiveBench web-site at http://bioinfo.pl/LiveBench to see an up-to-date evaluation of their performances.

REFERENCES

1. Fischer D, Baker D, Moult J. We need both computer models and experiments (correspondence). Nature 2001;409:558.
2. CAFASP1. Critical assessment of fully automated protein struc-

ture prediction methods. Proteins, Special Issue, 1999. See http://www.cs.bgu.ac.il/~dfischer/cafasp1/cafasp1.html.
3. CASP3. Critical Assessment of Protein Structure Prediction Methods (CASP), Round III. Proteins, 1999. Special Issue; see also http://Prediction center.lnl.gov.
4. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. LiveBench-1: continuous benchmarking of protein structure prediction servers. Protein Sci 2001;10:352–361.
5. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acid Res 1997;25: 3389–3402.
6. Fischer D, Elofsson A, Rychlewski L. MaxSub: a measure of quality assessment of protein structure predictions. Bioinformatics 2000;6:776–785.
7. Kelley LA, MacCallum RM, Sternberg MJE. Enhanced genome annotation using structural profiles in the program 3d-pssm. J Mol Biol 2000;299:501–522.
8. MacCallum RM, Kelley LA, Sternberg MJE. Structure assignment with text description-enhanced detection of remote homologues with automated swiss-prot annotation comparisons. Bioinformatics 2000;16:125–129.
9. Fischer D. Combining sequence derived properties with evolutionary information. Proc Pacific Symp on Biocomputing Jan 2000;119–130.
10. Fischer D, Eisenberg D. Protein fold recognition using sequence-derived predictions. Protein Sci 1996;5:947–955.
11. Jones DT. Genthreader: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
12. Rychlewski L, Jaroszewski L, Li W, Godzik A. Comparison of sequence profiles: strategies for structural predictions using sequence information. Protein Sci 2000;9:232–241.
13. Shi J, Blundell TL, Mizuguchi K. Fugue: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 2001;310: 243–257.
14. Mizuguchi K, Deane CM, Blundell TL, Overington JP. Homstrad: a database of protein structure alignments for homologous families. Protein Sci 1998;7:2469–2471.
15. Meller J, Elber R. Linear optimization and a double statistical filter for protein threading protocols. Proteins 2001;45:241–261.
16. Ogata K, Umeyama H. Prediction of protein side-chain conformations by principal component analysis for fixed main-chain atoms. Protein Eng 1997;10:353–359.
17. Ogata K, Umeyama H. An automatic homology modeling method consisting of database searches and simulated annealing. J Mol Graph Model 2000;18:258–272,305–306.
18. Johnson MS, Overington JP. A structural basis for sequence comparisons: an evaluation of scoring methodologies. J Mol Biol 1993;233:716–738.
19. Bates PA, Sternberg MJ. Model building by comparison at casp3: using expert knowledge and computer automation. Proteins 1999; 37:47–54.
20. McGregor MJ, Islam SA, Sternberg MJE. Analysis of the relationship between sidechain conformation and secondary structure in globular proteins. J Mol Biol 1987;198:295–310.
21. Bower MJ, Cohen FE, Dunbrack RL, Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J Mol Biol 1997;267:1268–1282.
22. Fariselli P, Casadio R. A neural network based predictor of residue contacts in proteins. Protein Eng 1999;12:15–21.
23. Pazos F, Olmea O, Valencia A. A graphical interface for correlated mutations and other structure prediction methods. CABIOS 1997; 13:319–321.
24. Ouzounis C, Perez-Irratxeta C, Saner C, Valencia A. Are binding residues conserved? In Fifth Annual Pacific Symposium on Biocomputing, Hawaii, World Scientific; 1998.
25. Goebel U, Sander C, Schneider R, Valencia A. Correlated mutations and contact in proteins. Proteins 1994;18:309–317.
26. Olmea O, Valencia A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Folding Design 1997;2:S25–S32.
27. Chew LP, Kedem K, Elber R. Unit-vector rms (urms) as a tool to analyze molecular dynamics trajectories. Proteins 1999;37:554–564.
28. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
29. Abagyan RA, Batalov S. Do aligned sequences share the same fold? J Mol Biol 1977;273:355–368.
30. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 1998;281:565–577.
31. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. J Mol Biol 1997;268:209–225.
32. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
33. Fischer D, Eisenberg D. Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. Proc Natl Acad Sci USA 1997;94:11929–11934.
34. Fischer D, Eisenberg D. Predicting structures for genome sequences. Curr Opin Struc Biol 1999;9:208–211.
35. Abbott A. Computer modellers seek out ten most wanted proteins (news). Nature 2001;409:4.
36. Fischer D, Elofsson A, Rychlewski L. The 2000 olympic games of protein structure prediction. Protein Eng 2000;13:667–670.
37. Cristobal S, Zemla A, Fischer D, Rychlewski L, Elotsson A. A study of quality measures for protein threading models. BMC Bioinformatics 2001;2:5.
38. Lundström J, Rychlewski L, Bujnicki J, Elotsson A. Pcons: A neural-network-based consensus predictor that improves fold recognition. Protein Sci 2001;10:2354–2362.