PROTEINS: Structure, Function, and Genetics 53:503–516 (2003)

CAFASP3: The Third Critical Assessment of Fully Automated Structure Prediction Methods

Daniel Fischer,[1]* Leszek Rychlewski,[2] Roland L. Dunbrack, Jr.,[3] Angel R. Ortiz,[4] and Arne Elofsson[5]
[1]*Bioinformatics, Department of Computer Science, Ben Gurion University, Beer-Sheva, Israel*
[2]*BioInfoBank Institute, Poznan, Poland*
[3]*Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, Pennsylvania*
[4]*Department of Physiology and Biophysics, Mount Sinai School of Medicine of New York University, New York, New York and Unidad de Bioinformatica, Centro de Biologia Molecular "Severo Ochoa" (CSIC-UAM), Universidad Autonoma de Madrid, Madrid, Spain*
[5]*Stockholmm Bioinformatics Center, Stockholm University, Stockholm, Sweden*

ABSTRACT **We present the results of the fully automated CAFASP3 experiment, which was carried out in parallel with CASP5, using the same set of prediction targets. CAFASP participation is restricted to fully automatic structure prediction servers. The servers' performance is evaluated by using previously announced, objective, reproducible and fully automated evaluation methods. More than 60 servers participated in CAFASP3, covering all categories of structure prediction. As in the previous CAFASP2 experiment, it was possible to identify a group of 5–10 top performing independent servers. This group of top performing independent servers produced relatively accurate models for all the 32 "Homology Modeling" targets, and for up to 43% of the 30 "Fold Recognition" targets. One of the most important results of CAFASP3 was the realization of the value of all the independent servers as a group, as evidenced by the superior performance of "meta-predictors" (defined here as predictors that make use of the output of other CAFASP servers). The performance of the best automated meta-predictors was roughly 30% higher than that of the best independent server. More significantly, the performance of the best automated meta-predictors was comparable with that of the best 5–10 human CASP predictors. This result shows that significant progress has been achieved in automatic structure prediction and has important implications to the prospects of automated structure modeling in the context of structural genomics. Proteins 2003;53:503–516.**

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

INTRODUCTION

Automatic structure prediction has witnessed significant progress during the last few years. A large number of fully automated servers, covering various aspects of structure prediction, are currently available to the community. A number of evaluation experiments aimed at assessing the capabilities and limitations of the servers exist.[1,2] These experiments help non-expert predictors to make better use of the automated tools because they provide predictors with valuable information that can help them in choosing which programs to use and in evaluating the reliability of the programs when applied to their specific prediction targets. One of these experiments is CAFASP,[3,4] where the participants are fully automatic Web servers, covering various aspects of protein structure prediction, and where the evaluation is conducted over automatically produced models without the human-expert intervention allowed at CASP. CAFASP is run in parallel with CASP, using the same set of prediction targets. This provides a unique opportunity to compare the performance of individual servers with that of humans. The parallel setting provided by CASP and CAFASP is also very useful for tool developers because directions for further improvements in the automated methods can be identified.

MATERIALS AND METHODS

The way CAFASP operates has previously been described[3,4] and has been announced in the CAFASP Web page at http://www.cs.bgu.ac.il/~dfischer/CAFASP3. All the methodology, predictions, and evaluation results are available at this site. We refer the readers to the site for details. In what follows we present only a brief outline.

Each model submitted to CAFASP underwent two independent evaluations: one conducted by the CAFASP automated methods, and the second, by the CASP human assessors. Here we report the results of the automatic CAFASP evaluation; please see the corresponding assessment reports in this issue for the server evaluation conducted by the CASP human assessors. The fully automated CAFASP evaluation is a unique feature of this experiment because only objective and reproducible evaluation methods are used. In addition, the details of the evaluation procedures were announced before the experi-

*Correspondence to: Daniel Fischer, Bioinformatics, Department of Computer Science, Ben Gurion University, Beer-Sheva 80415, Israel. E-mail: dfischer@cs.bgu.ac.il

Received 23 February 2003; Accepted 16 June 2003

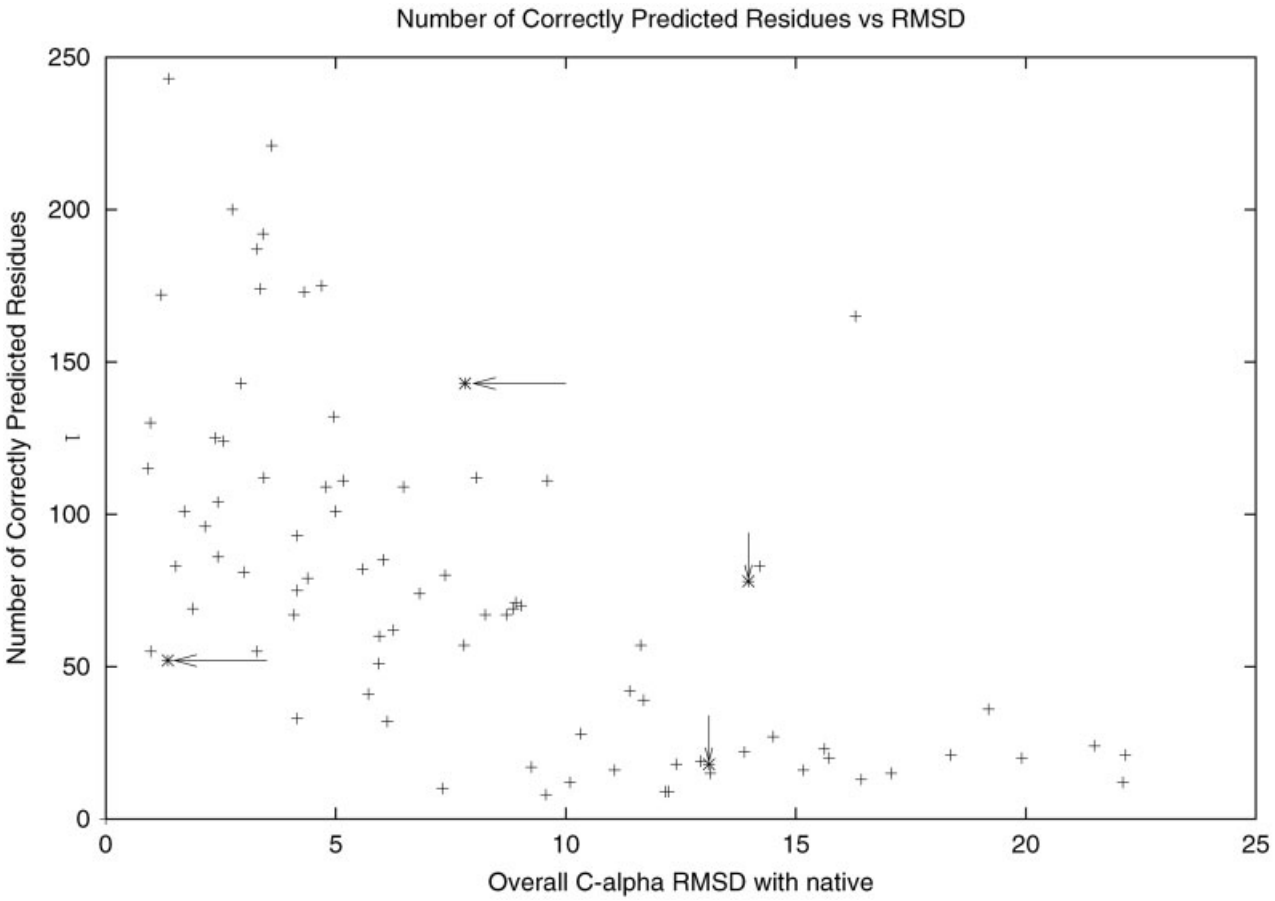

Fig. 1. Illustration of the MaxSub scoring system. Each dot in the figure corresponds to the rank-1 prediction selected by the 3D-JURY meta-predictor[21] for each of 85 CAFASP targets. The *x* axes correspond to the overall C-α RMSD of each model with its corresponding native structure. **a:** The reason why RMSD cannot be used as a measure to score partially correct models is shown. Models with relatively small RMSD to native contain a significant number of correctly predicted residues (i.e., superimposable over the native structure with a distance $<$ 3.5 Å; *y* axis). However, above an RMSD of $\approx$5 Å, both incorrect and partially correct models can obtain similar RMSD-s. For example, the two vertical arrows correspond to a partially correct model with 78 correctly predicted residues and an RMSD of 13.97 Å, and to a completely wrong model with only 18 "correctly predicted residues" with a similar RMSD of 13.11 Å. Thus, to be able to distinguish between partially correct models from the incorrect ones, one has to take into consideration the subset of "well-predicted" residues of each model. This is what is computed by MaxSub (and by a number of other measures such as GDT, lgscore,[8] MAMMOTH,[22] and the "S" measure used in LiveBench-6). A simple scoring procedure would be to add the number of well-predicted residues each model has. This is the procedure applied in the LiveBench-6 evaluation (see the LiveBench-6 report in this issue). **b:** MaxSub's normalization scoring system. MaxSub's scores (*y* axis) 1) normalize the size of the superimposable subset by dividing by the size of the target (thus accounting for the completeness of the prediction and avoiding the unequal weight that larger targets may have) 2) take into account the quality of the subset of superimposable residues, and 3) require a minimum size of superimposable residues to produce a positive score (thus avoiding the accumulation of low scores from incorrect predictions). For example, an excellent model containing 52 superimposable residues for a target of size 56 receives a high MaxSub score of 0.93, whereas a model with 143 superimposable residues for a target of size 252 receives a score of 0.39 (horizontal arrows). The models highlighted in (a) (vertical arrows) receive MaxSub scores of 0.37 and 0.00, respectively. Notice also that for RMSD values $<$ 5 Å, there is a good correlation between the MaxSub scores and RMSD, whereas RMSD's $>$ 5 Å, MaxSub effectively distinguishes between incorrect and partially correct models. This figure shows that most MaxSub scores $>$ 0.5 correspond to models with RMSD's to native $<$ 5 Å. From (a) and (b), it is clear that most of the models with zero MaxSub scores correspond to incorrect models (with only a small number of well-predicted residues) and that positive MaxSub scores correspond to models that can be considered to be at least partially correct.

ment began at the CAFASP3 Web site, so all participants knew how they would be evaluated.

The central CAFASP processing site was at http://bioinfo.pl.[5] On release of each target sequence, the servers had 48 h to automatically file their predictions. All predictions were immediately made available at the CAFASP Web site, making it possible for "meta-predictors" to analyze, use, and possibly improve the servers' results when filing their predictions. Thus, the challenge for meta-predictors was to produce more accurate models than those produced by the servers. Two types of meta-predictors should be distinguished. The first type corresponded to CAFASP automated meta-predictors, which had 48 additional h to file their meta-predictions. The other group corresponded to CASP human meta-predictors, which had access to the results of both the individual servers and of the automated meta-predictors, and had weeks before the CASP submission deadline. Because of the advantage that automated meta-predictors had, a direct comparison of the predictions filed by the servers

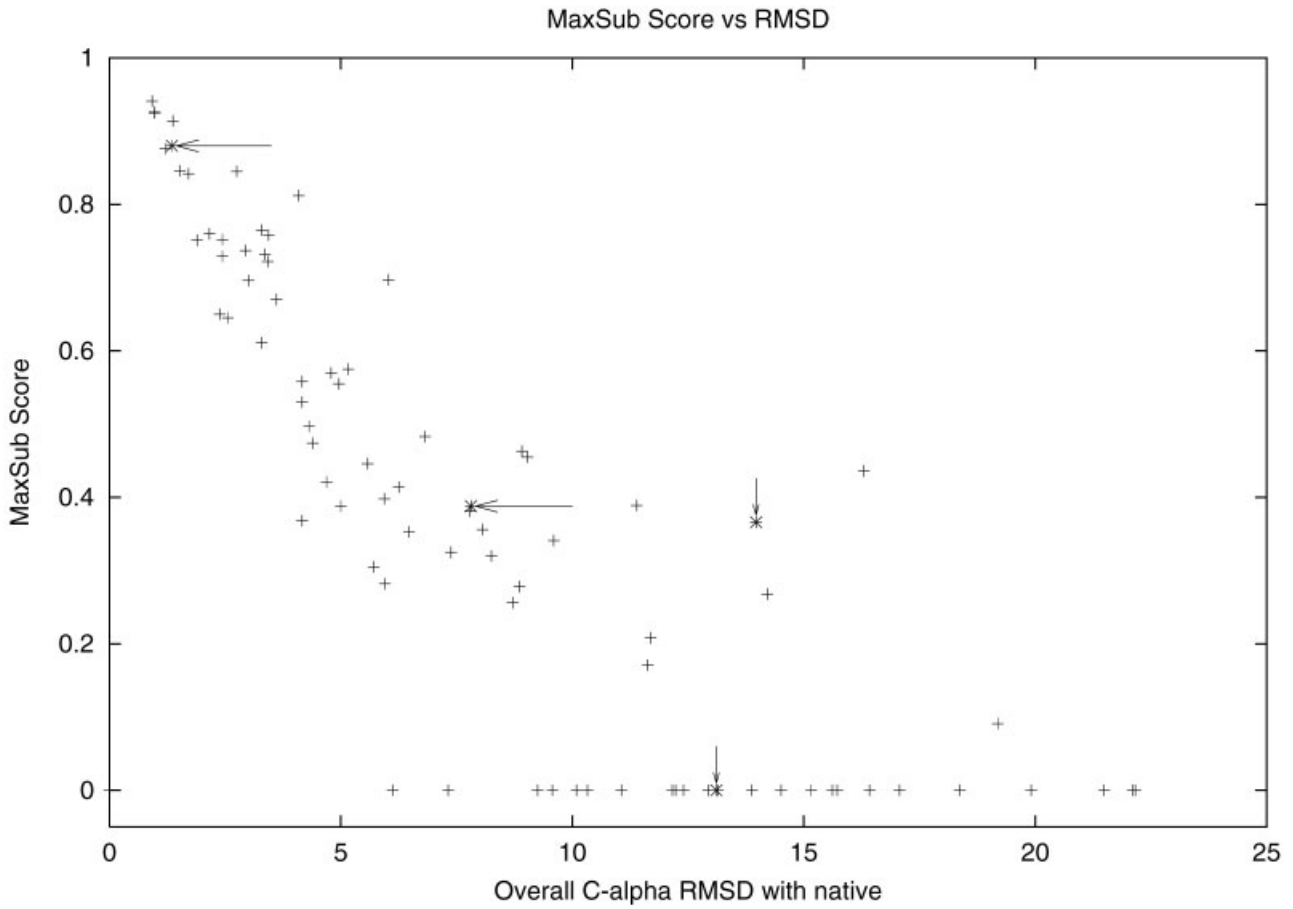

Figure 1. (Continued)

versus those filed by the meta-predictors is unfair for the former. However, a comparison of the performance of the human meta-predictors versus that of the automated meta-predictors (despite the fact that human meta-predictors had additional data and time) enables us to provide an indication of the value of human-expert intervention.

More then 60 servers, covering the five prediction categories of structure prediction registered at CAFASP3 (for a full list of participating servers, their full names, and the abbreviations used here, see the CAFASP3 Web site). Here we concentrate on the Homology Modeling, Fold Recognition, and "New Fold" (previously known as "ab initio") categories. The evaluation results of the Secondary Structure and Contact Prediction servers will be presented elsewhere by the corresponding CAFASP coordinators of these categories: Burkhard Rost and Alfonso Valencia.

For each target, up to 10 alternative models were allowed (corresponding to the top 10 ranks of the servers' output). The score a server received for a target was the score for the top rank only, but all the models were stored and made available.

In CAFASP3, the CASP5 targets were divided into two main divisions: Homology Modeling (HM: 32 targets) and Fold Recognition (FR: 30 targets). HM targets correspond to targets for which PSI-BLAST[6] found good matches (e-values < 0.001) to proteins of known structure. FR targets correspond to all the others. All targets were used as queries for all servers. Each division was evaluated separately. At the time each target was released, if it became clear that it contains more than one domain, we considered each separate domain as a separate target, in addition to the full target sequence. Please see the CAFASP3 Web site for a list of the domains used and for further details of the evaluation procedures.

RESULTS

We present separate results for each of the CAFASP categories.

Fold Recognition Targets

The scoring function used in CAFASP3′s FR evaluation was MaxSub,[7] a sequence-dependent assessment measure. MaxSub's performance as an evaluation measure has been assessed[7,8] and has been used as the scoring function in previous evaluation experiments.[2,4] Consequently, here we only describe it briefly. MaxSub attempts to identify the largest subset of superimposable C_α atoms of a model and an experimental structure (superimposable is defined to mean that the maximum distance after superposition between corresponding atoms is below the default value of

TABLE I. FR Sensitivity Results for the Independent Servers

N-1 rank	Servers	Score range	No. correct
4	raptor	3.98	13
5	shgu	3.93	13
6	orfeus	3.64	12
8	orf_c fugue3	3.39–3.67	11–12
10	fugsa orf_b	3.44–3.63	10–12
12	ffas03	3.36	11
13	inbgu arby	3.04–3.32	11
14	3dpssm	3.31	11
15	samt02	3.21	10
19	mGENTHREADER	2.84	9
...			
48	pdbblast	0	0

N-1 rank. The rank the server achieved throughout the application of our N-1 rule, described in the text. The application of the N-1 rule can rank a number of servers at the same rank, and not all ranks may be populated with servers. See also Table III.
Score range, the cumulative MaxSub score over all targets; No. correct, total number of correct (positive MaxSub score) predictions.

3.5 Å). MaxSub produces a single normalized score in the range 0.0 (for an incorrect prediction) to 1.0 (a perfect one). A MaxSub score above zero was considered to be a correct prediction. Figure 1 illustrates why a measure other than RMSD is required and provides some guidance as to the relative meaning of the MaxSub scores in comparison to RMSD.

The following is a summary of the results presented at the Asilomar meeting, which are also identical to those published at the CAFASP Web site. The structures of a handful of targets have been released since the completion of our analysis. Because the number of CASP targets is relatively small and the performance differences of many servers are slight, small changes in the relative rankings would occur, if the newly released structures were included in the analysis (results not shown).

Sensitivity results

Sensitivity was computed as the sum of the MaxSub scores of the rank-1 models over all targets. To rank the different servers, we applied the "N-1" rule, where N denotes the number of targets. This rule computes the rank that each server would obtain in each of the N = 30 subsets of size N-1. For each server, the best rank achieved in any of the N subsets was registered and is reported. The ranking was computed by considering all participating servers, but in what follows, we present separate results for the "independent" servers and for the automated "meta-predictors." Independent servers are defined as servers that do not use the output of other CAFASP3 servers. CAFASP3 meta-predictors are defined as servers that use the input of other CAFASP3 servers. Detailed evaluation results, including comprehensive tables, are available at our main Web page.

Table I lists the sensitivity performance of the top independent servers assessed over the 30 FR targets. The top performing servers, raptor[9] and shgu,[10] produced correct models for 43% of the FR targets (13 of 30). Their cumulative MaxSub scores were 3.98 and 3.93, with overall N-1 ranks of 4 and 5, respectively (for the missing ranks see Table III). Raptor is a new threading program using linear programming. Shgu is an independent version of the 3D-SHOTGUN algorithm,[10] which runs by using the internal components of the bioinbgu server[11] (the servers are referred to here with the same abbreviations as those listed in the CAFASP Web site; please see the site for their full names). Following raptor and shgu are a group of 10 servers producing between 10 and 12 correct models. These 10 servers are as follows: three variants of Rychlewski's orfeus method, orfeus, orf_c, orf_b (unpublished); two versions of Mizuguchi's fugue method: fugu3 and fugsa[12]; Godzik's new version of ffas: ffas03[13]; Fischer's inbgu[11] server; Lengauer's arby internal meta-predictor system (unpublished); Sternberg's 3dpssm server[14]; and Karplus's new version of the sam method, samt02.[15] Their cumulative MaxSub scores range from 3.04 to 3.64, with overall N-1 ranks of 6–15. Table I shows that many CAFASP servers are able to produce more correct predictions than the standard sequence comparison tool PSI-BLAST (locally run at bioinfo.pl under the name pdbblast), which ranks at the bottom of the table. It is clear that the differences among the top performing servers are not very large; thus, the precise ranks are likely to differ if the evaluation was conducted by using a larger number of targets or a different evaluation method. Nevertheless, a similar set of best performing servers has been identified by the large-scale LiveBench evaluation experiments, albeit with slight differences in their relative rankings (see Refs.[2,16, and 17] and Table I of the LiveBench-6 report in this issue).

We have analyzed the distribution of correct FR predictions by using the FR(H), FR(A), and NF CASP target classification. FR(H) includes targets having (distant) homologues of known structure, FR(A) includes targets having previously observed folds but without evidence of evolutionary relationship, and the NF class includes targets corresponding to novel previously unseen folds. Most of the correct server predictions corresponded to the proper FR targets [CM/FR, FR(H) and FR(A)]. Excellent to good rank-1 models were produced for all CM/FR targets and for all but one of the targets in each of the FR(H) and FR(A) classes. Among those targets with no (NF) or little (NF/FR) structural similarity to known structures, the servers had much lower success. No server produced any useful models at rank-1 for three of the targets in the NF/FR class nor for five of the targets in the NF class. This means that of the 30 FR targets considered here, more or less correct rank-1 models were obtained for at most 20 of the targets. However, as Table I shows, no individual server had correct models for more than 13 targets. This means that even the top performing servers failed to produce correct models for many targets where other servers succeeded.

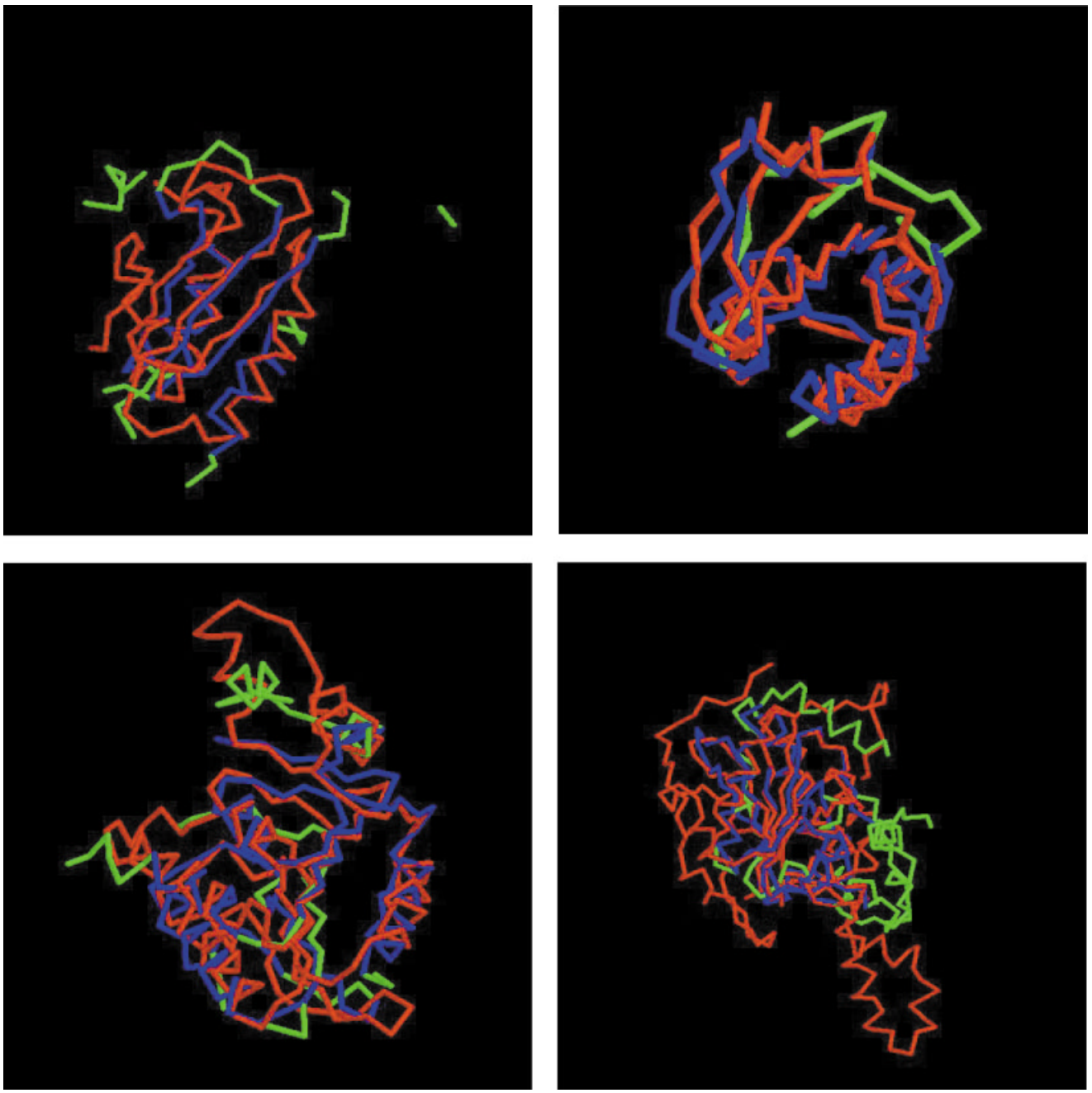

Fig. 2. Examples of outstanding independent server predictions. For clarity, we report the MaxSub score times 10. All targets shown here and in Figure 3 correspond to FR targets, as classified in the CAFASP3 Web site, except for target T0189 which was classified as HM-hard. **a:** Supfampp's model no. 6 on target T0130; 53 of 100 well-predicted residues with a MaxSub score of 4.1. **b:** 3dpssm's model no. 2 on target T0134; 87 of 106 well-predicted residues with a very high MaxSub score of 6.7. **c:** Raptor's model no. 9 on target T0136_1; 118 of 144 well-predicted residues with a MaxSub score of 3.7. **d:** mGENTHREADER's model no. 2 on target T0136_2; 121 of 205 well-predicted residues with a MaxSub score of 3.7. **e:** Prospect's model no. 2 on target T0159; 75 of 211 well-predicted residues with a MaxSub score of 3.6. **f:** Protab's outstanding model no. 8 on the NF/FR target T0170; 49 of 69 well-predicted residues with a MaxSub score of 5.6. **g:** Shgu's model no. 6 on target T0174_1; 69 of 197 well-predicted residues with a MaxSub score of 2.5. **h:** Pspt's model no. 2 on target T0174_2; 62 of 155 well-predicted residues with a MaxSub score of 3.0.

The main evaluation above considered rank-1 models only. Consequently, good predictions obtained by the servers at ranks below 1 cannot be appreciated. Examples of such independent servers' predictions are shown in Figure 2. These server predictions are among the best predictions for the corresponding targets and include predictions where none of the rank-1 server models were correct (e.g., the hard target T0174). The examples illustrate that many servers (not necessarily the ones ranked at the top in Table I) are often able to produce good predictions at higher ranks; thus, for the hardest targets, valuable results can be obtained when analyzing their top 5–10 predictions (see below).

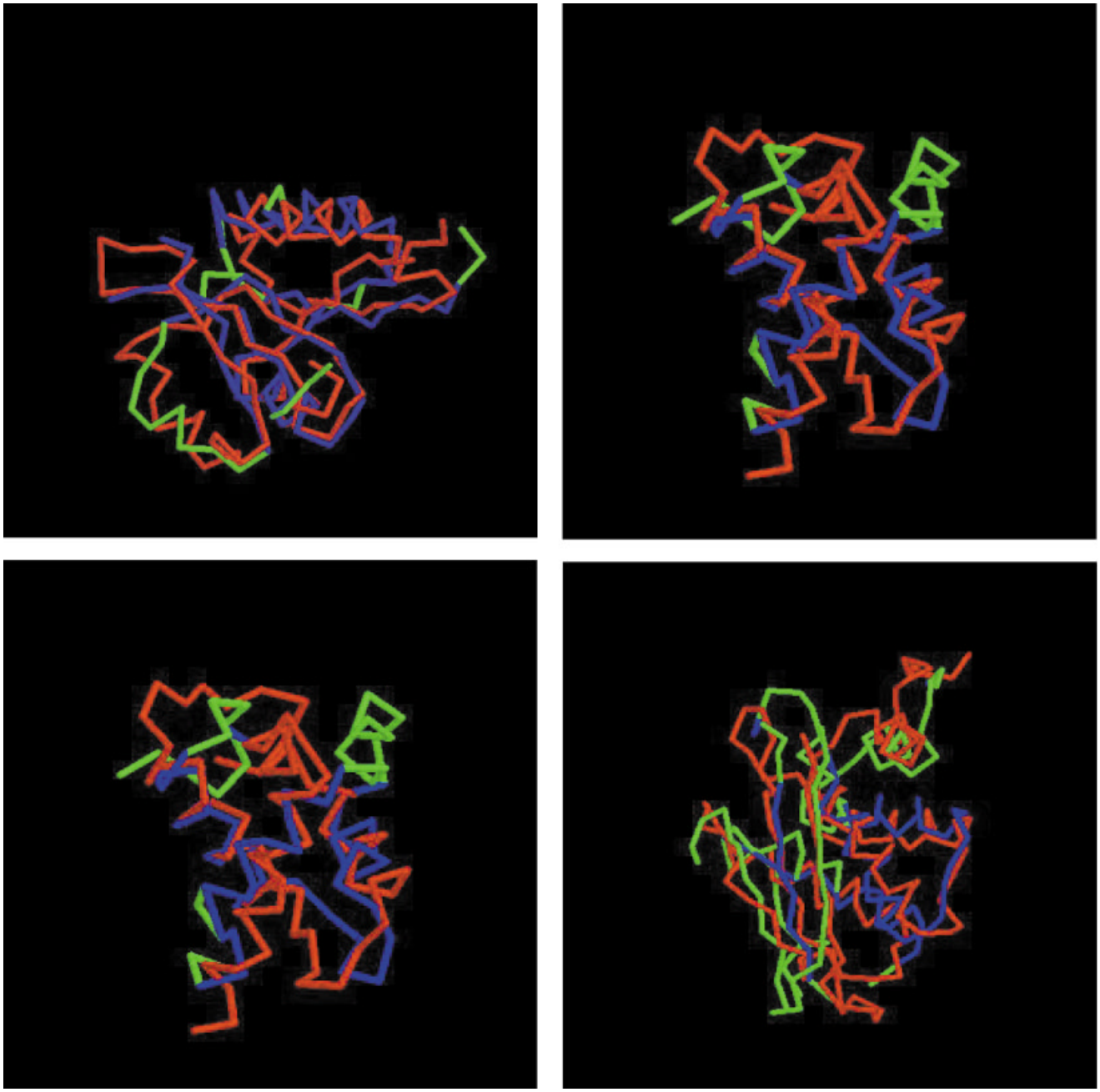

Figure 2. (Continued)

Specificity results

The value of any prediction program critically depends on the ability to attach confidence scores to each prediction. Thus, deriving confidence thresholds for prediction programs is essential for their automatic application. Servers usually assign confidence scores to their top hits. Thus, the evaluation experiments such as CAFASP and LiveBench, analyze the reliability of the servers in assigning scores to their predictions and derive recommended confidence thresholds at various error rate levels.

In CAFASP3, we computed specificity in a LiveBench-like way. For each server that had a parsable reliability score, we computed the number of correct predictions that it produced with scores better than those of its first, second, and third incorrect ones. Automated servers produce a single reliability score per model, regardless of the domain structure of the target. To obtain a meaningful specificity evaluation, we had to consider a different set of targets from that used in the sensitivity evaluation. For our specificity evaluation, we did not partition the targets into domains and considered only FR and "HM-hard" targets that had no domains in the "HM-easy" category (see the CAFASP Web site for the HM-hard and HM-easy classification and for the list of targets used in the specific-

TABLE II. Specificity Results for the Independent Servers

Server	Average specificity	Incorrect 1st		Incorrect 2nd		Incorrect 3rd	
shgu	21.4	19	22.87	20	15.92	21	14.18
inbgu	19.8	17	14.10	19	12.50	20	11.90
fugue3	19.0	12	8.78	17	4.88	21	4.55
ffas03	18.4	11	−20.40	15	−14.50	20	−8.30
orfeus	18.2	13	9.82	15	7.22	20	5.88
fugsa	18.2	11	8.78	17	4.88	20	4.55
raptor	17.8	11	10.85	17	7.57	18	7.45
3dpssm	17.8	13	0.07	15	0.11	18	0.24
orf_c	17.6	14	9.94	16	8.18	17	7.33
mGENTHREADER	17.4	6	0.70	18	0.56	20	0.53
orf_b	16.8	11	9.71	15	7.19	16	7.10
...							
pdbblast	13.0	11	0.02	12	0.03	13	0.04

1st, 2nd and 3rd: For each server, the number of correct predictions with scores better than those of its first, second, and third incorrect ones are listed. The first number shown corresponds to the number of correct predictions with scores better than the score for the corresponding incorrect prediction. The second number corresponds to the score reported by the server.

ity computation). For multidomain targets, the best domain result was considered. This procedure resulted in the selection of 33 targets to be used in the specificity computation. As in LiveBench, for each server, an average specificity was computed. Further details about the specificity computation can be found at the CAFASP web site.

Table II lists the most specific 11 independent servers. The table lists the number of correct predictions (from the selected set of 33 targets) with scores better than those of the first, second, and third incorrect ones. The three most specific servers were shgu, inbgu, and fugue3, with up to 64% (21 of 33) correct predictions before their third incorrect one. Most of the servers shown in Table II are also identified as the most sensitive ones (Table I), albeit with a different relative ranking. As with the sensitivity results, the most specific servers perform significantly better than the control server pdbblast. It is of interest that the specificity analysis of the LiveBench experiments identifies a similar set of most specific servers, but again with some differences in their relative ranking. Notice that here, as in the sensitivity evaluation, the differences among the most specific servers are only slight; thus, different evaluation methods on different sets of targets can result in a permutation of the ranks.

The magnitude of the scores of the first three false positives helps 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 shgu top hit with a score below the range 15–20 or a pdbblast top hit with a score above the range 0.02–0.04.

Are the servers as a group useful?

From previous experiments, it has become clear that often a correct prediction can be obtained by one server but not by the others. It has also been observed that no server can reliably distinguish between weak hits (predictions with below-threshold scores) and wrong hits and that often a correct model is found among the top hits of the server, but scoring below a number of incorrect models. From this and other observations, many human expert predictors have realized that to produce better predictions, the results from a number of independent methods need to be analyzed.

To study whether it was possible to obtain a better prediction using a very simple consensus method that used the information from several servers, in CAFASP2, we (the CAFASP organizers) filed predictions to CASP using a consensus of the servers' predictions named CAFASP-CONSENSUS.[4] The CAFASP-CONSENSUS ranked better than any of the individual servers, with only six other human CASP predictors ranking higher.[4] This finding illustrated the utility of the servers' results as a group. Following CAFASP2, Elofsson implemented the CAFASP-CONSENSUS ideas into the automated program Pcons.[18] Pcons receives as input the top models produced by three to eight different servers and selects the models that are evaluated to be more likely to be correct, based on the structural similarities among the input models. Pcons corroborated the strength of the consensus idea in the subsequent LiveBench experiments.

Since then, meta-prediction has become the most successful approach and has been applied by a large number of human predictors, including some of the best CASP performers (see their reports in this issue). Because manual meta-predicting can be a time-consuming task and because there are very many different ways to use the information of various servers to produce a meta-prediction, the development of automated meta-predictors has proliferated. The reports of some of these automated meta-predictors also appear in this issue, including those of the Pmodeller/Pcons and the 3D-SHOTGUN meta-predictor series, and the robetta and 3D-JURY meta-meta-predictors. The Pmodeller/Pcons series are variations on the Pcons theme; the Pmodeller meta-predictors apply the ProQ structural evaluation step[19] using the Pcons output as input to Sali's Modeller program[20] and thus generate full-atom models. The 3D-SHOTGUN series, developed by Fischer,[10] includes the 3dsn, 3ds3, and 3ds5 variations. 3D-SHOTGUN does not only select the most likely models from its input but it also assembles hybrid models, which on average are more complete and accurate than the individual models are. The robetta server (unpublished), developed by Baker, also receives Pcons output as input, which is further processed by using their rosetta method. The 3D-JURY system, developed by Rychlewski,[21] is another variant of the Pcons/3D-SHOTGUN meta-predictors and selects among the input models the ones more likely to correspond to a correct prediction.

In what follows we report the performance of the automated meta-predictors in CAFASP. In a separate section below, we compare the performances of automated versus human meta-predictors.

Table III lists the sensitivities of the CAFASP meta-predictors. The N-1 rank column indicates the rank ob-

TABLE III. FR Sensitivity Results for the Meta-Predictor Servers

N-1 rank	Meta-predictors	Score range	No. correct
1	3ds5 robetta	5.17–5.25	15–17
3	Pmodeller 3ds3 Pmodeller3	4.21–4.36	13–14
6	3dsn	3.90	13
7	Pcons3	3.75	12
13	libellula	2.96	10

See Table I.

tained by the meta-predictors when considering all the CAFASP servers together (independent and metas), so that a comparison with the data presented in Table I can be conducted. Such a comparison clearly shows that the performance of the top meta-predictors is significantly superior to that of the independent servers. The two most sensitive meta-predictors, 3ds5 and robetta, produced correct models for 50–56% (15–17 of 30) of the FR targets, followed closely by the Pmodeller, 3ds3, and Pmodeller3 meta-predictors. The best meta-predictors had overall MaxSub scores 30% higher than that of the best independent servers. Similarly, consistent results have also been observed in the LiveBench experiments, confirming the superior performance of these meta-servers. Examples of excellent predictions of the meta-predictors are shown in Figure 3.

Table IV shows that the specificities of the CAFASP meta-predictors are very similar, with the best of them having specificities 15–25% higher than those of the independent servers (Table II). Although the most specific independent server, shgu, had 21 correct predictions before its third incorrect one, the 3ds5 meta-predictor had 25. The scores for the first incorrect predictions listed in Table IV are also consistent with those identified in the LiveBench experiment and thus, should be a useful indication for the recommended confidence thresholds for the users of the meta-predictors.

Homology Modeling Targets

For the N = 32 HM targets we have applied the exact same evaluation as the one used for the FR targets. However, for brevity, we present the HM results without separating the independent servers from the meta-predictors.

Table V lists the HM sensitivity performance of the top CAFASP independent and meta-predictor servers. Most servers succeeded in producing correct models for all the 32 HM targets. Consequently, only the cumulative MaxSub scores are listed in the table. The top performing independent servers here were orf_b, samt02, shgu, inbgu, fugue3, and 3dpssm, with only slight differences in their cumulative MaxSub scores and with many others closely following. The 3ds3 and 3ds5 meta-predictors perform slightly above the best independent servers, but the other meta-predictors score as well as many independent servers. The control pdbblast server performs significantly below the top servers and very similarly to the two homology-modeling servers esypred and famsd.

The 32 HM targets include 20 HM-easy targets (for which the first iteration of PSI-BLAST hits a PDB entry) plus 12 HM-hard targets (for which templates are identified only at subsequent PSI-BLAST iterations). The latter set includes a number of relatively challenging prediction targets, with relatively low sequence similarities. Many of the models for the HM-hard targets contained significant alignment errors. Because some servers may be particularly aimed at the traditional HM targets, where the alignment problem is not a major issue, we analyzed the performance of the servers on the HM-easy and the HM-hard targets separately. The results (not shown, but full tables are available at the CAFASP Web site) not unexpectedly show that meta-servers performed significantly better at the HM-hard targets. It is of interest that in the HM-easy targets, the best performers were three independent servers at rank-1: samt02, orf_b, and inbgu, with the homology-modeling server esypred at rank 3. Thus, it seems that there is no (MaxSub detectable) added value from the meta-predictors among the HM-easy targets.

It is clear that our evaluation does not find large differences between the top servers, probably because the MaxSub evaluation focuses on the accuracy of the alignments using only C_α atoms. Further differences may be detected if the details of the loops and the correctness of the side-chains (for those servers generating full-atom models) were assessed. Roland Dunbrack made such an analysis, and the rest of this section contains his independent report (further details are available at http://www.fccc.edu/research/labs/dunbrack/cafasp3.html).

Only servers producing full-atom models for 27 of the easier targets were assessed. The sequence identities between targets and templates range between 8 and 43%, and a total of 270 predictions were analyzed. To evaluate side-chain prediction accuracy, we used percent χ_1 correct within 40°. This criterion is used almost universally in presentation of new side-chain prediction methods, alongside RMS, χ_1, and volume overlap measures. Looking at all the predictions as a whole, the average prediction accuracy can be expressed roughly as $P = I + 25$, where P is the percent correct χ_1 and I is the sequence identity. However, the range in accuracy on any one target was as high as 30%. In correctly aligned regions, as defined by MaxSub, the accuracy results were about 5% higher.

Because each model for a given target does not contain the same number of predicted residues, we can look at side-chain accuracy either with the total number of correctly predicted χ_1 dihedrals or as a percentage correct of the predicted side-chains for each server. Looking at total number of correct side-chains, robetta performed best with 2741 correct followed by famsd and protinfocm with 2476 and 2445, respectively. The number of correct side-chains for the remaining servers ranged from 1929 (pcomb) to 2388 (fams). Judging by the correctly aligned regions alone, the ranking is nearly the same, except that protinfocm drops from third to sixth. If we judge the servers by percent correct of the number of side-chains in each server's models, then robetta still has the highest percent-

age with 56.5% correct. Protinfocm has 55.3% correct. Alax, fams, and famsd range from 51.3 to 52.9%, and the remaining four servers (pcomb, esypred, jigsaw, and Pmodeller) cluster in a range of 44.7–47.7%. Although the accurate prediction rate of the top two groups is within the 95% confidence interval (±1.5%), robetta produced longer models, for a significantly higher total of correct side-chains. Protinfocm performs slightly better in structurally aligned regions, with 64.6% compared to robetta with 62.7%. The remaining servers again cluster in two groups: alax, fams, and famsd with 58.2–59.8% correct and pcomb, Pmodeller, and esypred with 48.0–50.4% correct.

We also assessed local backbone accuracy as the percentage of residues with both backbone ϕ and ψ dihedral angles correct within 30°. This is in lieu of a more detailed analysis of particular loop predictions. Loop predictions are quite difficult to assess, because it is not always clear what alignment and parent structure were used for each prediction. In an overall assessment, the percent correct ϕ, ψ can be expressed roughly as $P = (4/3)I + 40$ (I is the percent sequence identity). Any target with 40% sequence identity would be expected to have P near 100%. Judging by the number of correct residues on these 27 targets, the best servers were robetta (3612) and Pmodeller (3563). The remaining servers (famsd, fams, esypred, alax, pcomb, jigsaw, and protinfocm) ranged between 2880 (protinfocm) and 3339 (famsd). Compared to the side-chain results, the biggest difference is that Pmodeller performs very well in this measure but has relatively poor side-chain performance. Protinfocm performs poorly in the backbone measure but well in the side-chain measure. The other groups are ranked roughly the same by both measures. As a percentage of predicted residues, Pmodeller ranks first with 75.6% followed by robetta with 74.5% correct (95% confidence interval is ±1.2%). Protinfocm is last with 65.2%. The other six groups fall between 69.1 and 72.1%.

Human Versus Machine Comparison

CASP5 provides a unique opportunity to compare the performance of servers with predictions by manual experts. As stated above, such a comparison is unfair, as is the comparison of the independent servers with the meta-predictors, because the predictions of the servers were available to meta and human predictors. An additional advantage for human CASP participants was access to the automated meta-predictor results (plus additional weeks to file the predictions). Thus, it is not unexpected that some human CASP predictors clearly outperformed the individual servers. However, a rather surprising result was obtained when the human CASP performance was compared with that of the automated meta-predictors. In a way, all human CASP participants are meta-meta-predictors, in the sense that they could (but not all necessarily did) use the information of the automated methods published in the CAFASP Web site. The following is based on the evaluation of all the automated servers and the human participants, which has independently been prepared by Michael Levitt, available at http://csb.stanford.edu/levitt/CASP5_AutoAssessor. This evaluation is based on the official GDT_TS scores published at the CASP Web site and using the same target categories as the human assessors. The ranks are based on the sum of z scores above 1.0. For further details please see the above Web site.

Table VI lists the ranks of the top three human predictors and of the top automated servers. The automated meta-predictors, robetta, 3ds3, and 3ds5, as in the CAFASP evaluation, perform better than the best individual servers and were ranked among the top 20 groups using the sum of GDT_TS z scores ranking. The best performing individual server is shgu at rank 23, which corresponds to an independent server (non-meta) using the same 3D-SHOTGUN algorithm as 3ds3 and 3ds5 do, but applied internally to the five components of the bioinbgu method. Following closely are the Pmodeller and Pmodeller3 meta-predictors. The performance of the best manual groups was outstanding; the best human z score was more than twice larger than that of the best automated server, indicating, as in CAFASP2, that the best human predictor at CASP clearly did much better than the best server. However, with the exception of about 10 human groups, the performance of the best automatic methods is comparable to that of many human groups. In addition, we notice that a vast majority of the groups that ranked at the top 20 ranks corresponded to human groups that either also had a server in CAFASP or that extensively used the CAFASP results to prepare their predictions (see the corresponding reports in this Issue).

A more detailed analysis can be obtained by separating the targets into the CASP CM and FR classes (we do not include here the separate analysis of new fold category because it contained a very small number of targets, composed of a number of domains; see below). As the difficulty of the targets increases, the difference in performance between humans and servers seems to decrease. In the CM class, the best meta-predictor, 3ds3, is at rank 9, with the other meta-predictors and shgu ranking below rank 26. This ranking is very similar to the one obtained in the overall analysis, probably because of the abundance of CM targets in the experiment. However, the best human z score here was less than twice as large as that of the best server.

In the FR class, a slight increase in the relative ranks of the automated servers is observed, with robetta at rank 6 and other meta-predictors below rank 20. The best human z score here was only 66% larger than that of the best server. The fact that humans appear to be better at the easier targets may indicate that with relatively correct sequence-structure alignments, humans, throughout the application of a number of verification tools, are able to produce better refined models than those produced by the automated servers.

One certain advantage that human groups have over the current automatic methods is that the former can often identify multidomain targets, whereas most automatic servers usually return a prediction for one domain only. A similar analysis to the one presented above (not shown), in which the best prediction for any domain of a target

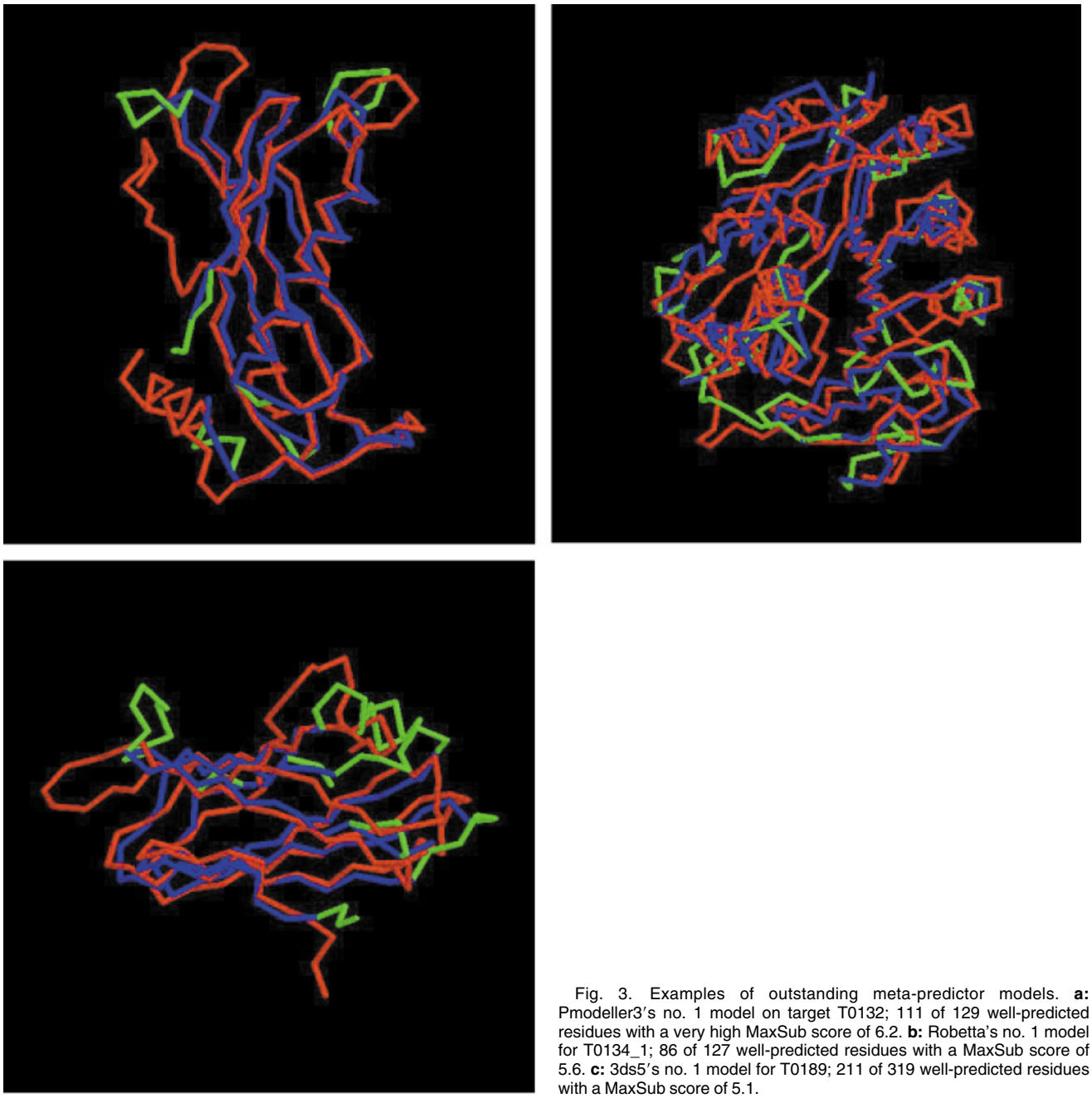

Fig. 3. Examples of outstanding meta-predictor models. **a:** Pmodeller3's no. 1 model on target T0132; 111 of 129 well-predicted residues with a very high MaxSub score of 6.2. **b:** Robetta's no. 1 model for T0134_1; 86 of 127 well-predicted residues with a MaxSub score of 5.6. **c:** 3ds5's no. 1 model for T0189; 211 of 319 well-predicted residues with a MaxSub score of 5.1.

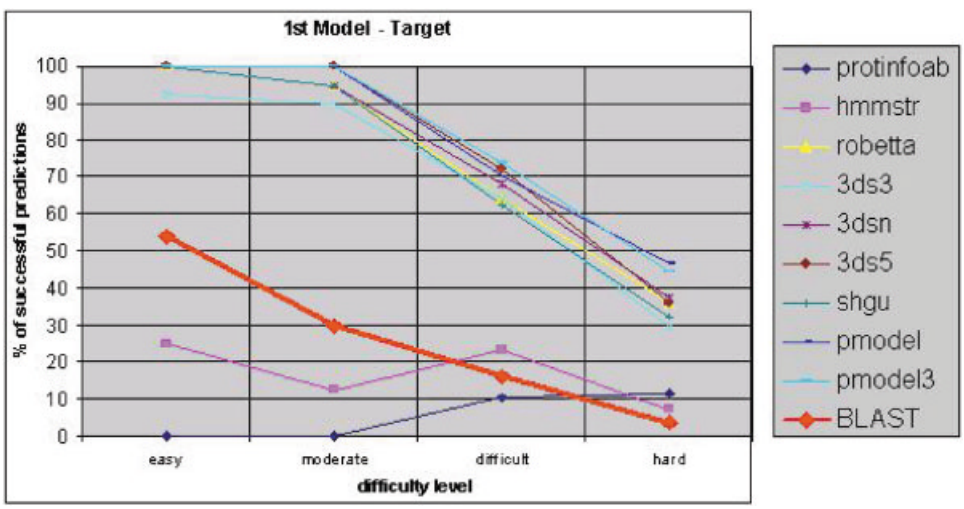

Fig. 4. MAMMOTH analysis of the performance of selected servers at the different target difficulty levels (see text). PDB-BLAST is shown in red as reference.

TABLE IV. Specificity results for the Meta-Predictor Servers

Server	Average specificity	Incorrect 1st		Incorrect 2nd		Incorrect 3rd	
3ds5	24.8	20	4.99	24	4.17	25	4.06
Pmodeller	22.6	20	1.34	21	1.33	23	1.25
3dsn	22.2	19	30.32	21	27.12	22	25.97
3ds3	22.2	19	30.32	21	27.12	22	25.97
Pmodeller3	22.0	16	2.10	22	1.70	23	1.66
Pcons3	21.6	17	1.79	19	1.57	23	1.34

See Table II.

TABLE V. Sensitivity Results on the HM Targets†

N-1 rank	Servers	Score range
1	3ds3	19.89
2	3ds5	19.79
3	**orf_b samt02 shgu**	19.27–19.52
4	Pmodeller	19.37
5	Pcons3	19.25
6	**inbgu fugue3**	19.06–19.07
7	3dsn robetta **3dpssm** Pmodeller 3	18.70–19.04
8	**orfeus**	19.06
14	**orfblast ffas03 famsd**	18.34–18.75
15	**pdbblast**	18.56
17	**esypred**	18.24

See Table I.
†Independent servers are shown in bold.

protein was selected, shows that this slightly increased the relative servers' ranking. However, this analysis also showed that successful domain handling can only partially explain the superior human performance. Another advantage of human predictors is that they may be better at selecting the best templates.

Another interesting observation is that in CASP4 the semiautomated meta-predictor CAFASP-CONSENSUS ranked seven overall. In CASP5, the best meta-predictors ranked within the top 10 ranks. Thus, the relative difference in performance between humans and machines has not changed significantly. If we assume that the current meta-predictors perform better than the CAFASP-CONSENSUS of CASP4, then we could conclude that both humans and automated methods in CASP5 have improved since CASP4. Although we cannot prove it, this assumption appears to be reasonable because the first implementation of Pcons performs similarly to the CAFASP-CONSENSUS[18] and current meta-predictors have been evaluated to be superior to Pcons.[2,17]

New Fold Category

Because the MaxSub evaluation presented above is relatively stringent, mainly focused on HM and FR targets, most of the models for the hardest targets obtained a zero MaxSub score (using default parameters). To assess the performance of the servers on these hardest targets, our new fold coordinator, Angel Ortiz and his colleagues, have applied a separate evaluation. A summary of their main findings is presented in the following paragraphs.

All models from CAFASP servers that directly produce C_α coordinates were reevaluated. These include the models of the ab initio servers Protinfoab and Hmmstr, those of the meta-predictors robetta, 3ds3, 3dsn, 3ds5, Pmodeller, and Pmodeller3 and those from the independent server shgu. This reevaluation was conducted by using the evaluation method named MAMMOTH.[22] MAMMOTH is a structural alignment program. The algorithm first finds structural correspondences between two proteins and then attempts to find a core of residues with small RMSD between both structures. A similarity score is then computed from the number of residues found in this core. The final MAMMOTH score for the alignment is given by $s = -\ln P$, where P is 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 of the scores resulting from random structural alignments. A MAMMOTH score of 5.0 corresponds to a probability of a random match of $\approx$0.001. Studies in our laboratory indicate that a score of $\approx$5 is expected for structure pairs related at the fold level in SCOP and that this usually corresponds to a situation where 20–25% of the residues are within 4.0 Å RMSD in the alignment. We refer the readers to Ref. [22] for further details.

Models were partitioned into domains according to the domain definitions provided in the CASP5 Web page. Each target domain was scanned against the SCOP database with MAMMOTH to select the nearest structural neighbor or optimal template for the target. The target was then assigned to the SCOP family of the template if the MAMMOTH score was 5.0. Otherwise, the target was declared to be a new fold. Finally, according to the target-template score, targets were assigned to four difficulty levels: easy, moderate, difficult, and hard.

Each target and the top-ranking (first) model were compared with MAMMOTH. Models scoring 5.0 were considered successful predictions. The server success at the different difficulty levels was then measured by the number of successful predictions of a server normalized by the total number of predictions produced by that server at that difficulty level.

Figure 4, shows our main results. For the meta-predictors, three features are worthnoting: 1) A linear decrease in performance, with a steep slope, in going from moderate to hard targets for all servers, 2) large improve-

TABLE VI. Comparison of Human and Server Ranking

Rank	Human/server	Name	*Z* score
All: CM + FR + new fold			
1	Human	Ginalski	75.94
2	Human	Bujnicki-Janusz	59.00
3	Human	Baker	57.76
...			
10	Server (meta)	robetta	36.14
13	Server (meta)	3ds5	34.62
16	Server (meta)	3ds3	32.39
23	Server	**shgu**	26.85
27	Server (meta)	Pmodeller	24.85
30	Server (meta)	Pmodeller3	23.37
CM only			
1	Human	Bujnicki-Janusz	47.17
2	Human	Ginalski	46.55
3	Human	GeneSilico	42.82
...			
9	Server (meta)	3ds3	25.60
15	Server (meta)	3ds5	22.16
21	Server (meta)	robetta	18.46
22	Server	**shgu**	17.30
26	Server (meta)	Pmodeller	15.84
FR only			
1	Human	Ginalski	24.26
2	Human	Skolnick-Kolinski	21.64
3	Human	Baker	19.55
...			
6	Server (meta)	robetta	14.56
13	Server (meta)	Pmodeller3	9.00
15	Server (meta)	3ds5	8.09
16	Server	**arby**	7.86
17	Server (meta)	3ds3	7.29
22	Server	**shgu**	6.72

Human/Server. Meta-predictor servers are indicated with the word "meta" in parentheses. "Human" corresponds to human predictors that filed their predictions to CASP and not to CAFASP. Independent (i.e, not meta-predictors) servers are shown in bold.
Z-score. Sum of *z* scores for top CASP human and CAFASP servers as listed in Michael Levitt's independent evaluation (see http://csb.stanford.edu/levitt/CASP5_AutoAssessor). Two significant differences are observed in the server rankings of this table compared with those of the CAFASP evaluation. The first is that the independent server arby ranks very high, and the second is that raptor did not reach the top ranks. We have no explanation for this, but it could probably be attributed (at least in part) to differences in the evaluation methods and/or to format differences between the CASP and CAFASP stored models.

ments over PDB-BLAST at all but the hard level, with significant improvements at the moderate level of difficulty, where the meta-predictors show a sustained high performance. This corresponds mainly to remote relationships at the family level according to SCOP (not shown). 3) A strikingly similar behavior is observed for all servers in this category. For purely ab initio methods, the most important result is a lack of overall predictive power at all levels. A final result to note is an equally unsatisfactory prediction ability of PDB-BLAST for difficult and hard targets. This is somewhat expected in this particular case, because targets in the new fold category were selected for being undetectable by BLAST. Nevertheless, we included the BLAST performance to provide a suitable baseline to compare the servers.

DISCUSSION

The CAFASP3 results show that automated servers are able to produce excellent models for all the 32 HM targets and for about half of the 30 FR targets. Among the 10 hardest FR targets (which include up to 8 NF targets), no useful rank-1 models were produced by the servers. In all categories, many servers perform significantly better than the widely used PSI-BLAST program.

Because the number of targets in CAFASP is relatively small and because the differences among the independent servers is small, it is difficult to establish an exact ranking of the servers. However, from a number of different evaluations, it can be safe to conclude that a group of about 10 servers correspond to the best performers. These include raptor, shgu, orfeus, orf_c, orf_b, fugue3, fugsa,

ffas03, inbgu, arby, 3dpssm, and samt02. Each of these servers failed to produce correct models for many targets where other servers succeeded. The above general observations are consistent with the results of the previous CAFASP and LiveBench experiments.

One of the main clear results of CAFASP3 is that the servers as a group provide very valuable information that can be exploited by both human and automated meta-predictors. The performance of the best automated meta-predictors is ≈30% higher than that of the best independent servers. Another clear result was that although about 10 human predictors outperformed the automated servers, the performance of the best automated meta-predictors is outstanding. The superior performance of all meta-predictors critically depends on the availability of the independent servers; without the latter, the former could not exist. In addition, the superior performance of human and automated meta-predictors is mainly a success of all the independent servers as a group, suggesting that very valuable information is produced by each of them, not only in their rank-1 model. This is further suggested by the fact that meta-servers perform relatively simple processes, mainly identifying structural recurrences among the input models.

Thus, automated meta-predictors represent a significant advance in the field. Meta-predictors have attracted some criticism, being characterized as "exploiters" of the independent servers and as dangerous tools that may preclude the availability and development of more servers. However, it seems that meta-prediction has become a standard practice for most human predictors. Because this practice is time-consuming and error-prone, if conducted by hand, the need to automate the meta-prediction process is obvious. This is what meta-predictors have achieved, relieving human expert predictors from the tedious tasks of gathering the results from the independent servers and analyzing them to produce a meta-prediction.

In principle, a meta-predictor could operate as a single, independent program that comprises all the individual components. However, this seems to be difficult to imple ment, mainly because structure prediction usually involves a long series of individual, independent components (PSI-BLAST searches, secondary structure prediction, template recognition, refinement modules, etc.). In addition, because the methods used by the different servers are sometimes conflicting, exploiting different aspects of protein sequence and structure, it is not obvious how they can all be combined into a single method.

An important aspect of meta-predictors is to go beyond the simple selection of a model out of their input (i.e., to have some added value). The Pmodeller series is a valuable contribution because it adds to the fold recognition process, the ProQ structural analysis step, and the full-atom-generating step using Modeller. The 3D-SHOTGUN series assembles hybrid accurate and more complete models than the input models. Finally, the robetta meta-predictor only uses Pcons' results when the latter are highly reliable; in the other cases, the ab initio rosetta program is used to produce a model. Thus, all these meta-predictors provide additional information, not directly available from the input models.

Despite the encouraging results of CAFASP3, it is important to note that there is still a long way to go. First, very few correct predictions were produced for the hardest targets. Second, many of the predictions that were considered here to be correct, in fact, are only partially correct, containing significant incorrectly modeled regions. Third, automated servers perform poorly for multidomain targets; automatic domain identification and partitioning is clearly one important aspect of future improvements. Fourth, even among the easiest HM targets, many models still require significant refinement before they can reach accuracies comparable to those of medium- or low-resolution experimental structures. Fifth, full-atom refinement, side-chain placing, and loop modeling continue to be bottlenecks.

One aspect of CAFASP that needs improvement is in the automatic evaluation methods themselves. Alternative evaluations of the CAFASP data have been conducted, generally confirming the main results presented here, but with a considerable reshuffling of the relative server rankings (links to some of these alternative evaluations are available at the CAFASP Web site). Other evaluation procedures are being investigated in the context of the LiveBench experiment. One of the 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. To better assess the capabilities of servers, large-scale experiments, such as LiveBench and EVA,[23] are required. These provide an important complement to the CASP/CAFASP experiments and, together, provide a better picture of the capabilities in the field.[1,2]

As an extension of the CAFASP and LiveBench experiments, a number of new continuous experiments have been recently created[2] (links to the Web sites of these new experiments appear in the CAFASP Web page). One of them is the PDB-CAFASP experiment, which like LiveBench, is the continuous, large-scale version of CAFASP. The difference is that the targets used in PDB-CAFASP are "on-hold" PDB entries, whereas in LiveBench, the targets are newly released entries. Thus, PDB-CAFASP allows for the gathering of blind predictions that can be evaluated when the corresponding structures are released. With use of the data from PDB-CAFASP, two other experiments are now available. The first is MR-CAFASP, which is aimed at evaluating the value of predicted models in the structure determination process itself. The second experiment is LSD-CAFASP, which is aimed at evaluating the modeling capabilities of side-chains and loops.

Having a clear assessment of what automated structure prediction can and cannot do is critical for their large-scale applicability (e.g., in structural genomics[24]). Paradoxically, as the number of known structures increases, so increases the number of sequences that biologists expect to model. Only if automated, reliable tools are able to model most of the proteins closely and distantly related to

proteins of known structures will the goals of structural genomics be achieved.

ACKNOWLEDGMENTS

CAFASP was made possible throughout the voluntary efforts of the CAFASP organizers and category coordinators, including B. Rost and A. Valencia. The results of CAFASP and CASP would be very different without the availability of the predictions from the independent servers. We thank the developers of the independent servers for making their predictions available to the human and automated meta-predictors; the developers of the automated meta-predictors for making their meta-predictions available to the human CASP participants; the CASP organizers, assessors, and predictors for encouragement; the many users of our servers and the experimentalists who provided the targets. Thanks also to Yaniv Azaria for his help in preparing the CAFASP Web pages and to Dmitry Lupyan and Alejandra Leo for their generous help with the MAMMOTH analysis presented here.

REFERENCES

1. Fischer D, Elofsson A, Rychlewski L. The 2000 olympic games of protein structure prediction. Protein Eng 2000; 13:667–670.
2. Fischer D, Rychlewski L. The 2002 olympic games of protein structure prediction. Protein Eng 2003; 16:157–160.
3. CAFASP1. Critical assessment of fully automated protein structure prediction methods. Proteins 1999; Special issue. See http://www.cs.bgu.ac.il/~dfischer/cafasp1/cafasp1.html.
4. Fischer D, Elofsson A, Pazos F, Valencia A, Rost B, Ortiz AR, Dunbrack RL Jr. CAFASP2: the second critical assessment of fully automated structure prediction methods. Proteins 2001; Suppl 5:171–183. Special Issue; see also http://www.cs.bgu.ac.il/~dfischer/CAFASP2.
5. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. Structure prediction meta server. Bioinformatics 2001; 17:750–751.
6. Altschu 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.
7. Siew N, Elofsson A, Rychlewski L, Fischer D. MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 2000; 16:776–785.
8. Cristobal S, Zemla A, Fischer D, Rychlewski L, Elofsson A. A study of quality measures for protein threading models. BMC Bioinformatics 2001; 2:5.
9. Xu J, Ming L, Lin G, Kim D, Xu Y. Protein threading by linear programming. Proceedings Pacific Symposium on Biocomputing, 2003. Forthcoming.
10. Fischer D. 3D-SHOTGUN: a novel, cooperative, fold recognition meta-predictor. Proteins 2003; 51:434–441.
11. Fischer D. Combining sequence derived properties with evolutionary information. Proceedings Pacific Symposium on Biocomputing, Jan 2000. p 119–130.
12. 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.
13. 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.
14. Kelley LA, MacCallum RM, Sternberg MJE. Recognition of remote protein homologies using three-dimensional information to generate a position specific scoring matrix in the program 3D-PSSM. *RECOMB 99, Proceedings of the Third Annual Conference on Computational Molecular Biology,* 1999. p 218–225.
15. Karplus K, Barrett C, Hughey R. Hidden markov models for detecting remote protein homologies. *Bioinformatics* 1998;14:846–856.
16. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. Livebench-1: continuous benchmarking of protein structure prediction servers. *Protein Sci* 2001;10:352–361.
17. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. LiveBench-2: large-scale automated evaluation of protein structure prediction servers. *Proteins* 2001; Suppl 5:184–191. see also http://bioinfo.pl/LiveBench.
18. Lundstrom J, Rychlewski L, Bujnicki J, Elofsson A. Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci* 2001;10:2354–2362.
19. Wallner B, Elofsson A. Can correct protein models be identified? *Protein Sci* 2003;12:1073–1086.
20. Sali A, Blundell T. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 1993;234:779–815.
21. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* 2003;19:1015–1018.
22. Ortiz AR, Strauss CE, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci* 2002;11:2606–2621.
23. Eyrich VA, Marti-Renom MA, Madhusudhan MS, Fiser A, Pazos F, Valencia A, Sali A, Rost B. EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics* 2001; 17:1242–1243.
24. Fischer D, Baker D, Moult J. We need both computer models and experiments (correspondence). *Nature* 2001;409–558.