PREDICTIONS FROM AUTOMATIC SERVERS

CAFASP-1: Critical Assessment of Fully Automated Structure Prediction Methods

Daniel Fischer,[1]* Christian Barret,[2] Kevin Bryson,[3] Arne Elofsson,[4] Adam Godzik,[5] David Jones,[3] Kevin J. Karplus,[2] Lawrence A. Kelley,[6] Robert M. MacCallum,[6] Krzysztof Pawowski,[5] Burkhard Rost,[7] Leszek Rychlewski,[8] and Michael Sternberg[6]

[1]*Department of Mathematics and Computer Science, Ben Gurion University, Beer-Sheva, Israel*
[2]*Computer Engineering, University of California, Santa Cruz, California*
[3]*Department of Biological Sciences, University of Warwick, Coventry, United Kingdom*
[4]*Department of Biochemistry, Stockholm University, Stockholm, Sweden*
[5]*The Burnham Institute, La Jolla, California*
[6]*Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, London, England*
[7]*European Molecular Biology Laboratory, Heidelberg, Germany*
[8]*San Diego Supercomputer Center, La Jolla, California*

ABSTRACT **The results of the first Critical Assessment of Fully Automated Structure Prediction (CAFASP-1) are presented. The objective was to evaluate the success rates of fully automatic web servers for fold recognition which are available to the community. This study was based on the targets used in the third meeting on the Critical Assessment of Techniques for Protein Structure Prediction (CASP-3). However, unlike CASP-3, the study was not a blind trial, as it was held after the structures of the targets were known. The aim was to assess the performance of methods without the user intervention that several groups used in their CASP-3 submissions. Although it is clear that "human plus machine" predictions are superior to automated ones, this CAFASP-1 experiment is extremely valuable for users of our methods; 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 users in choosing which programs they wish to use and in evaluating the reliability of the programs when applied to their specific prediction targets. In addition, evaluation of fully automated methods is particularly important to assess their applicability at genomic scales.**

For each target, groups submitted the top-ranking folds generated from their servers. In CAFASP-1 we concentrated on fold-recognition web servers only and evaluated only recognition of the correct fold, and not, as in CASP-3, alignment accuracy. Although some performance differences appeared within each of the four target categories used here, overall, no single server has proved markedly superior to the others. The results showed that current fully automated fold recognition servers can often identify remote similarities when pairwise sequence search methods fail. Nevertheless, in only a few cases outside the family-level targets has the score of the top-ranking fold been significant enough to allow for a confident fully automated prediction. Because the goals, rules, and procedures of CAFASP-1 were different from those used at CASP-3, the results reported here are not comparable with those reported in CASP-3. Nevertheless, it is clear that current automated fold recognition methods can not yet compete with "human-expert plus machine" predictions.

Finally, CAFASP-1 has been useful in identifying the requirements for a future blind trial of automated served-based protein structure prediction. Proteins Suppl 1999;3:209–217. © 1999 Wiley-Liss, Inc.

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

INTRODUCTION

Fold recognition addresses a subset of the more general problem of prediction of the three-dimensional structure of a protein from its amino acid sequence. Fold-recognition methods search a library of known folds to find the most compatible protein for a given target sequence of unknown structure. The predictive power of such methods was clearly demonstrated in blind tests, such as Critical Assess-

Grant sponsor: GlaxoWellcome; Grant sponsor: National Institutes of Health; Grant number: GM60049; Grant sponsor: National Science Foundation; Grant numbers: BIR-9408579, DBI-9808077; Grant sponsor: DOE; Grant number: DE-FG03-95ER62112.

*Correspondence to: D. Fischer, Department of Mathematics and Computer Science, Faculty of Natural Science, Ben Gurion University, Beer-Sheva 84015, Israel. E-mail: dfischer@cs.bgu.ac.il

Received 2 February 1999; Accepted 28 May 1999

ment of Techniques for Protein Structure Prediction (CASP-3), where prediction targets were not known at the time the predictions were made. Groups developing structure prediction algorithms have made their methods available to their potential users: scientists interested in seeking structural insights for their particular research problems. Several fold-recognition methods are currently available to the community via web-servers. Assessing the performance of fully automated methods provides users with essential information about the methods' capabilities and limitations. Unfortunately, the CASP experiments have assessed the performance of "human plus machine" only, where the human part has usually been the expert developer of the method.

In this report, we demonstrate the performance of fully automated fold-recognition methods on a set of sequences, selected from the prediction targets of the CASP-3 experiment. Most of us took part in this meeting and submitted predictions, which in some cases are discussed in other reports in this special issue of *Proteins.* Therefore, it is important to stress the differences between results presented in this manuscript and those available on the official CASP-3 web site and discussed in other manuscripts in this issue.

Results presented here were generated automatically by web servers from submitted sequence data, with no human-expert intervention involved. In this respect, predictions presented here represent raw data that could be available to any user of a fold prediction server. An expert could easily improve such predictions, but his or her work often includes steps that are difficult to describe in a quantitative way and that are not easily reproducible. Such expertise was used extensively for predictions presented at the CASP-3 meeting. In many cases extensive work by groups of several people was necessary to prepare a successful prediction. Thus, it is not possible to measure from CASP-3 whether the success or failure of a group depended on the program used alone or on the human expertise applied. Someone lacking the exact specialized expertise in fold prediction may not be able to reproduce many of the predictions submitted at the CASP-3 meeting. On the other hand, results in this study are obtained automatically with no human intervention. Another difference between the results presented here and those from CASP-3 is that for the latter, all the predictions were blind, whereas in CAFASP-1 (which was carried out *after* the CASP-3 meeting), the experimental structures of all but four of the targets were already known. Thus, CAFASP-1 is not a blind experiment. In addition, because the goals of CASP-3 and CAFASP-1 are different, the evaluation procedures used were also different. Consequently, the results shown here are not comparable with those reported in CASP-3, nor should this article be regarded as CASP-3 compliant.

Our insistence on full automation of the fold-prediction process is not meant to belittle or to cast any doubt on the importance of the specialized expertise in fold predictions. But for certain purposes, such as the evaluation and comparison of various prediction algorithms and strategies, it is useful to have fully automated and easily reproducible methods. CAFASP-1 assesses the performance of the methods only and not the performance of humans using machines. This is what a non-expert user is most interested in: "Which program(s) should I choose to use in order to predict the structure of this new sequence, and how much can I trust it?" And last, but not least, fully automated methods are necessary to apply fold prediction to large groups of protein sequences, such as those available from genome projects. CAFASP-1 attempts to provide the wider community with an assessment of the capabilities and limitations of current fold-recognition servers.

METHODS

The Automated Methods

Seven groups actively participated in CAFASP-1. Table I lists for each group, in alphabetical order, the name of the server evaluated, its url, and the corresponding reference; Table II summarizes the main characteristics of the servers.

The Targets

For CAFASP-1 we have used as benchmark for our methods the CASP-3 targets. We have classified the targets into seven categories.

1. Targets with folds at SCOP's family level

In this category we chose the five targets with lowest sequence similarity to their corresponding folds from those classified in CASP-3 as having homologs at the family level according to SCOP[1]; the other family-level CASP-3 targets were excluded from CAFASP-1 because they do not pose any challenge to fold-recognition methods. The targets of this category included in CAFASP-1 are T0055, T0057, T0068, T0070, T0062 (for a full protein description, see the assessors' articles in this issue).

2. Targets with folds at SCOP's superfamily level

These targets are T0074, T0081, T0083, T0063, T0053, T0044, T0054, T0085, T0080.

3. Targets with folds at SCOP's fold level

These targets are T0046, T0071, T0043, T0067, T0059.

4. Multidomain targets

The domain boundaries for these targets were determined after the structures were known. Predictions for the single domains were included in CAFASP-1 to evaluate how performance on single domains changes. The domains considered are T0083.1, T0063.1, T0063.2, T0071.1, T0071.2, T0079.1, T0079.2 (the exact domain boundaries used are available from the CAFASP-1 main web page at http://www.cs.bgu.ac.il/~dfischer/cafasp1/cafasp1.html).

5. Targets of unknown fold

The structures of these targets have not yet been determined, and thus they correspond to genuine blind predictions: T0045, T0051, T0072, T0078.

TABLE I. The Groups Participating in CAFASP-1

Programs	url	Ref.	Comments
3D-PSSM and (1D + 3D)-PSSM	http://bonsai.lif.icnet.uk/foldfitnew/index.html	19	Two methods, both accessible from the same url
BASIC	http://cape6.scripps.edu/leszek/genome	23	The server is being moved to the following address: bioinformatics.burnham-inst.org
frsvr_SDP, frsvr_SDPM, frsvr_SDPMA2	http://www.doe-mbi.ucla.edu/people/frsvr/submit.html and http:/www.cs.bgu.ac.il/~bioinbgu	8	Three methods run for each submission
GenTHREADER	http://globin.bio.warwick.ac.uk/ psipred	4	
Karplus1, Karplus2, Karplus3	http://www.cse.ucsc.edu/research/compbio/HMM-apps/model-library-search.html or /T98-query.html or /model-library-search.html	11 20	Three variants of the SAM-T98 method
pscan	http://www.biokemi.su.se/~server/pscan/profscan2_ga11.html	21	
PSI-BLAST; BORK, NCBI	http://dove.embl-heidelberg.de/3D http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-psi_blast	17[a] 6	Two servers running supposedly the same program
topits	http://www.embl-heidelberg.de/predictprotein	22	

[a] The 3D server of the Bork group uses PSI-blast with an E-value threshold E = 0.001. It is thus reliable, but also rather conservative in its predictions; it does not report homologies below the threshold. Thus, its results are not comparable to those of the other servers that always report the top ranking hits, regardless of their scores. PSI-BLAST at NCBI does report below threshold hits, and like Bork's server, the only correct hits found were the five family-level targets plus target T0074.

6. Targets with new folds

These targets are T0052 and T0056. They were included in CAFASP-1 for additional evaluation purposes (see below).

7. Targets that could not be evaluated in CAFASP-1

We included in this category those CASP-3 targets for which we found that our evaluation procedure could not be applied (see below): T0061, T0075, T0077, T0079.

The Evaluation Method

In CAFASP-1 we were interested not only in fully automated predictions, but also in fully automated evaluation methods. We could not apply the exact evaluation procedures used in CASP-3 because 1) CASP-3 evaluation was not fully automated and 2) although many of the programs used for CASP-3 evaluation are extremely valuable, we could not apply them on time for CAFASP-1 results. Given these limitations and strict time constraints, we adopted the following (more limited) evaluation scheme. We expect that future CAFASP evaluations will more closely resemble those used in CASP.

In CAFASP-1, we used an evaluation scheme that tested only fold recognition, and not alignment quality. Each server produced a list of top-scoring folds, and if the first correct fold appeared at rank i, then 1/i points were awarded. (This scoring scheme is similar to that used for CASP-1 and other benchmarks.[2]) The rationale of this scoring system is as follows: Suppose a program always has the correct answer within the top i ranks; if only a single answer is desired, then, on average, the correct fold will be predicted with probability 1/i.

For structure prediction, evaluating the quality of the sequence-structure alignments is critical, since fold-recognition methods can in some cases produce poor sequence-structure alignments. Unfortunately, for CAFASP-1 evaluating alignments was not possible given the time constraints. Thus, our evaluation procedure may award points to predictions that in CASP-3 were considered incorrect. As progress in evaluation has been observed from CASP-1 to CASP-2 and CASP-3, we hope that for CAFASP similar progress will be observed and alignment quality will be properly assessed in CAFASP-2.

Another difficulty with this evaluation method is to identify what is the list of "correct" hits for each target. For the targets in category 1 (family-level) and 2 (superfamily-level) there was almost no difficulty. Any PDB[3] chain with the same fold type in SCOP as the target was considered a correct hit; anything outside the fold type of the target was considered to be wrong. The only exception was T0085, which belongs to a "fold" in SCOP, which, according to SCOP, is not a real fold, but only a collection of different folds. Thus, for T0085 we only accepted as correct hits those entries in T0085's superfamily.

Because we used SCOP to determine what folds were to be treated as correct, any reported hits that did not have a SCOP classification were excluded from the ranking before scoring.

In addition, we had to decide how to evaluate multidomain targets (T0063, T0081, T0083, and T0079), each of which has two domains. For the full-sequence tests, hits belonging to the fold type of either domain were considered correct. The separate domains of these targets were evaluated in the domain-level category (see below).

Unfortunately, we had to exclude T0079 from the full-sequence tests and include it only in the domain tests. As a whole, it was not easy to select a single fold type as correct for T0079; several fold types can be considered as good hits, but only for the individual domains.

For the targets in category 3 (fold-level) we applied the same criteria (accept as correct hits those entries classified in SCOP as the same fold type). However, there were three targets (T0077, T0061, and T0075) for which we could not determine what single fold type should be considered as a

TABLE II. Summary of the Servers' Characteristics

	Input: single sequence (SA) or multiple alignment (MA)	Uses predicted secondary structure Y/N	Sequence-Structure compatibility function	Algorithm to rank folds and to align	Gap regime	Fold library (updatable automatically (A) or with human intervention (H))	Conf. threshold	Other characteristics
3D-PSSM	SA	YES DSC (indirect use of multiple alignments)	Match to multiple sequence PSSM generated from 3D superposition of superfamily members + secondary structure	global	regular open/extend, end gaps penalized	H nonredundant 40% from SCOP	<0.4 is <5% errors per query	Born in 1998, preliminary version used at CASP3
(1D + 3D)-PSSM	SA	As 3D-PSSM	(i) 3D-PSSM results pooled with (ii) match to multiple sequence PSSM generated from sequences + secondary structure	As 3D-PSSM	as 3D-PSSM	as 3D-PSSM	as 3D-PSSM	as 3D-PSSM
BASIC	SA expanded to MA	NO	Profile to profile alignment (no structural information)	Local alignment	as in standard local alignment	H. Subset of PDB with 50% seq. id.	Experimental: E-value <1.0 and Z-score above 7.0	Experimental version. Last server update April 1998.
Gen THREADER	SA	NO	THREADER2 distance dependent pair potentials and solvation terms + neural net	Global	Open + Ext (no end penalties)	H, based on CATH	Conf.: above 0.8 = 1% error	n.a.
frsvr_SDP	SA	YES, PHD (indirect use of multiple alignments)	Weighted predicted secondary structure combined with Gonnet table	Global—local for ranking, local for aligning top ranks	Regular open/extend	H. Nonredundant at 50% seq. id. with multidomains cut also into single domains (>2,000 entries)	>7.00 very reliable, >5.00 interesting to look at	Born in 1996; results of CASP3 based on frsvr.
frsvr_SDPMA	MA, compiled by high hits with single BLAST search on SWISSPROT	YES, as SDP	Weighted predicted secondary structure (as SDP) combined with the average scores from the multiply aligned sequences	as SDP	as SDP	H, as SDP	Not yet assessed, but >7.00 is reliable, and as SDP	As SDP
frsvr_SDPMA2	MA, as SDPMA	YES, as SDP	As SDPMA	As SDP	As SDP plus special gaps for predicted or observed loops	H, as SDP	As SDPMA	Born in 1998; experimental version
Karplus1	SA	NO	HMM made from template sequence using SAM-T98	Local alignment sum-all-paths for ranking, global Viterbi for alignments	Column-dependent affine gaps	Manually updated, based on FSSP, PDBselect, and some hand-chosen others	n.a.	Uses only sequence info
Karplus2	Generates multiple alignment using SAM-T98 method	As Karplus1	HMM made from target sequence using SAM-T98	As Karplus1	As Karplus1	All of PDB, updated weekly	n.a.	As Karplus1
Karplus3	Combines Karplus1 and Karplus2	As Karplus1	Adds scores from Karplus1 and Karplus2	One of Karplus1 or Karplus2	As Karplus1	Combines Karplus1 and Karplus2	See error vs. score in ref. 12	As Karplus1
pscan	SA	N	A combined seq-str compatibility function using distance potentials as well as environment terms	Local	Regular open/extend	Fold Library from 1996 (346 entries)	Not tested	Born in 1996; not updated since
TOPITS	SA	YES (indirect use of alignment)	Predicted 1D structure (secondary structure, accessibility) combined with sequence information (McLachlan)	YES	Regular open/extend	H	zscore >4.5 => 90% zscore >3.0 =>60% correct first hits	Results FULLY automatic; unchanged since 1996; on WWW since 1995

TABLE III. Summary of the Results of CAFASP-1

	Family (5.00)[†]	Superfamily (9.00)	Fold (5.00)	Subtotal (19.00)	Domains (7.00)	Total (26.00)
3D-PSSM	4.50/90%	2.25/25%	1.25/25%	8.00/42%	2.00/29%	10.00/38%
(1D + 3D)-PSSM	5.00/100%	2.50/28%	1.33/27%	8.83/46%	3.81/54%	12.64/49%
BASIC	5.00/100%	3.12/35%	1.58/32%	9.70/51%	3.50/50%	13.20/51%
frsvr_SDP	5.00/100%	2.60/32%	0.67/13%	8.27/44%	2.83/57%	11.10/43%
frsvr_SDPMA	5.00/100%	3.48/39%	1.00/20%	9.48/50%	4.78/68%	14.26/55%
frsvr_SDPMA2	5.00/100%	4.43/49%	0.79/16%	10.22/54%	4.60/66%	14.82/57%
GenThreader	5.00/100%	5.00/56%	1.11/22%	11.11/58%	2.79/40%	13.90/53%
Karplus1	5.00/100%	2.72/30%	2.06/41%	9.78/51%	2.44/35%	12.22/47%
Karplus2	5.00/100%	3.24/36%	1.10/22%	9.34/49%	3.10/44%	12.44/48%
Karplus3	5.00/100%	3.52/39%	1.62/32%	10.14/53%	3.30/47%	13.44/52%
pscan	4.00/80%	1.67/19%	1.00/20%	6.67/35%	2.08/30%	8.75/34%
PSI-BLAST-NCI	5.00/100%	1.00/11%	0.00/0%	6.00/32%	—[a]	—
PSI-BLAST-BORK	5.00/100%	1.00/11%	0.00/0%	6.00/32%	—	—
TOPITS	3.50/70%	2.00/22%	1.33/27%	6.83/36%	—	—

[a] —, Domain results were not submitted for these servers.

[†] The numbers shown in parentheses represent the maximum attainable score in each column. The percentages shown after the "/" represent the percentage of the maximum score. A detailed table including the normalized scores can be seen on our summary results web page at http://www.cs.bgu.ac.il/~dfischer/cafasp1/results.html.

correct hit. In addition, the similarities of these targets to known folds are weaker and do not cover the full sequences. Thus, we decided not to evaluate these targets and place them together with T0079 into the category of non-evaluated targets (category number 7; the results of the automated methods on these targets are included in the CAFASP-1 web page). The targets in category 7 can be considered as not suitable for our simple evaluation method.

In category 4 we placed the domain predictions. Although determining the exact boundaries of a domain requires knowledge of the structure, we wanted to evaluate our methods also on single domains. In many cases, rough domain boundaries are also known from the sequence alone, as has been demonstrated in several cases in the CASP-3 predictions. The evaluation criteria used here was the same as for the above categories.

No evaluation was possible for category 5, as the structures of these targets are still unknown. Thus, our predictions for these targets are truly blind predictions. When the structures are released we will be able to evaluate them.

Finally, the two targets considered to be novel folds were placed in category 6. An ideal fold-recognition method should be able to identify also new folds, or at least give a very low score for the top-ranking fold. We filed predictions for targets in this category to observe how high our top-ranking fold scored for novel folds. This can be helpful in setting confidence thresholds for the methods when applied at larger scales (see below).

The lists of folds considered to be correct hits for each target are included in the CAFASP-1 web page.

Normalized scores

To identify some of the remarkable predictions, that is, good predictions for targets that few predictors did well on, we applied the following normalization. Let $S_{i,j}$ be the score received by program *i* on target *j* (computed as one divided by the rank of the first correct hit). Let T_j be the sum of scores for target *j*. We define the normalized score as $S_{i,j} * S_{i,j}/T_j$. The larger the normalized score, the more remarkable the prediction of program *i* for target *j* is.

The normalized scores did not change the overall rankings of the programs by much, and so are not shown, although the information is available on the main web page and the summary results web page accessible from it.

The grading system

To allow for a detailed comparison of performance we computed for each category and program a partial total of the inverse-rank scores. For each category we observed which programs obtained the highest total and subsequently added all the scores into an overall grade.

Reproducibility and Validity of the Automated Results

We verified that the results reported here accurately correspond to those that are obtained by the automated programs. Most results were checked by at least one other person (besides the developer of the program). Thus, a reader can submit the sequence of any of the targets and expect to obtain essentially the same results (excluding the differences that will appear due to possibly updated databases; as time passes, larger differences are likely to appear; because of the addition of new structures and/or intermediate sequences, some of the targets will become easier to predict). All the programs included in CAFASP-1 are available freely through the internet.

RESULTS

Table III is a summary of the results by program and by category. The individual inverse-rank scores by target and program are available in the corresponding tables from our main web page.

Category 1: Family-Level Targets

Almost all methods did well on the five close homology targets, although three methods had some trouble with target T0068 (see Table III and the results web page). To really distinguish between the performance of the methods in this category, evaluation of the alignments and scores is needed (see Discussion).

Category 2: Superfamily-Level Targets

In this category, the two best-scoring programs (GenThreader and SDPMA2) received 5.00 and 4.43 points, respectively (column "SUPERFAMILY" in the table), out of a possible maximum of 9.

GenThreader's performance is remarkable in that its 5.00 points were obtained from correct predictions at rank 1 in five targets (and zeroes in the other four); fsrvr_SDPMA2 had correct hits in the top ten ranks for eight of the nine targets, but only scored the correct hit first for three of the targets.

GenThreader[4] uses a combination of traditional profile-based sequence alignment, a set of threading potentials,[5] and a neural network to evaluate the quality of the implied structural model (see Table II). The profiles used here are generated using PSI-BLAST.[6] For each sequence-structure alignment, the threading potentials are summed for the implied model, and the energy sums and sequence alignment score are presented to a neural network. The neural network detects favorable combinations of energy sums and alignment scores and has been trained on known structural similarities found in the CATH structure classification database.[7] In the CASP-3 predictions made by the Jones group, GenThreader was only used as a pre-filter to detect superfamily matches. Apart from targets T0074, T0083, and T0085, the GenThreader results were not considered significant, and so most of the results were arrived at using a full threading method.

frsvr_SDPMA2 is a variation of the SDP method previously described,[8] which takes as input a multiple alignment of sequences homologous to the target and the predicted secondary structure given by PHD[9] (see Table II). The sequence-to-structure compatibility function combines 1) the sequence similarity between the multiple alignment and the sequence of a protein of known structure with 2) the extent of agreement between the predicted and the observed secondary structures. After CASP-3, folds of newer (or C_α only) PDB entries corresponding to the best matches of some targets were added to the fold library. For filing predictions for CASP-3, the Fischer group used human intervention, and in some cases, the fold obtained at rank 1 by frsvr was chosen.[10] In other cases, because the score of the rank 1 result was below a confidence threshold, a different fold was chosen.

The highest normalized scores in this category were also obtained by GenThreader and SDPMA2 for their correct prediction at rank 1 of target T0085 (shown in the CAFASP-1 web page). The second highest normalized score was obtained by three programs (BASIC, Karplus2, and Karplus3) in their correct identification at rank 1 for target T0044.

Category 3: Fold-Level Targets

The results for category 3 are shown in column "FOLD" of Table III. The best-performing methods were Karplus1 and Karplus3 with 2.06 and 1.62 points, respectively, out of a possible maximum of 5. Clearly, the performance of our methods in the "fold-level" targets is not as good as that in the "superfamily" targets. The most outstanding result when observing the normalized scores was obtained by Karplus1 on target T0043; it was the only program identifying the correct fold at rank 1.

The Karplus1 and Karplus3 methods are both SAM methods.[11] In SAM-T98 a hidden Markov model (HMM) is constructed from a single sequence and homologs that are found in a non-redundant protein database. The method alternates between searching the database for homologs using an HMM and realigning the homologs using Baum-Welch[11] training on the HMM. Only sequence information is used, not structure information. All scoring with HMMs was done with local scoring summing over all alignments. For Karplus1, an HMM was built for each fold in the fold library, and the target sequence scored against all the HMMs. For Karplus3, the HMM scores for the template and target methods were added. For CASP-3, hand-selection among the top few hits and hand-realignment was done, but subsequent analysis indicates that the fully automatic method does about as well overall as modification by hand.

It is interesting to notice that the best performers in this category were methods based on sequence-information alone. We have no explanation for this phenomenon, but, for example, previous tests of the SAM-T98 method indicated that it found the correct fold only when it found the correct superfamily. One possible partial explanation is that the SAM-T98 method relies on local alignment, so one does not need to match the entire fold to find a match. However, it is not possible to arrive at far-reaching conclusions from a sample of only five test cases.

When computing the subtotal from categories 1–3 the best performers are GenThreader, SDPMA2, and Karplus3 with 11.11, 10.22, and 10.14 points, respectively (column SUBTOTAL in Table III). These are the same top three as for the superfamily category alone, which contributes most of the variation between inverse-rank scores.

Category 4: Domain Targets

This category does not strictly belong to a fully automated context because determination of the domain boundaries required previous knowledge. Nevertheless, because in an actual prediction experiment it is often suspected what the boundaries are, we also tested our programs using the exact domain definitions. In this category the best performers were SDPMA and SDPMA2 with 4.78 and 4.60 points, respectively, out of a maximum of 7.00. The SDPMA methods identified the correct fold in rank 1 for four targets and in rank 2 for one target. The next best performer in this category was 1- and 3-dimensional position specific substitution matrix [(1D + 3D)-PSSM] with 3.81 points.

Fold recognition by 3D-PSSMs uses the SCOP database to identify remote homologues that are superposed in three-dimensions to obtain sequence alignments that could not be obtained from sequence alone. In addition, sequence-based profiles are generated to form 1D-PSSMs. The fold library consists of representative protein <40% identity from the SCOP. The search algorithm is a global dynamic programming algorithm with predicted secondary structure for the probe being matched with experimental secondary structure of the template. In 1D-3D-PSSMs, searches are made with the 1D- and the 3D-PSSMs, and the results are pooled and sorted by expectation E-value. The algorithms have been extensively developed since they were used at CASP-3.

The most remarkable normalized score achieved in this category was obtained by BASIC on target T0071.2. BASIC identified the correct fold at rank 2, whereas only two other methods had the correct fold at ranks 8 or higher. The next most remarkable normalized result was obtained by SDPMA and SDPMA2 for identifying the correct fold of target T0063.1 in rank 2, whereas only one other method had the correct fold at rank 9.

Clearly, knowing the exact domain boundaries of a target sequence contributes significantly to the performance of most methods. For example, for T0083, four methods did better when given the correct domain; for T0063, six methods did better and one worse on the domains; and for T0071, nine methods did better on the domains.

Total Scores Over all Categories

When summing up all scores from all four categories, the top programs are SDPMA2, SDPMA, GenThreader, and Karplus3 with 14.82, 14.26, 13.90, and 13.44 points, respectively. There is no evaluation for targets in categories 5–7, but the automated predictions submitted can be seen on our main web page.

Selectivity of the Methods

In addition to the sensitivity of the methods (i.e., the number of correct predictions), we have analyzed their selectivities. For a given threshold score *s*, we define selectivity here as the number of true-positives at rank 1 with scores better than *s* (in other words, the number of predictions in which the fold at rank 1 was the correct fold, and its score was better than *s*). To this end, for each method we compiled its rank 1 predictions, including the two targets with new folds (category 6). Then we set three threshold scores, Th1, Th2, and Th3 (different for each method), corresponding to the scores of the first, second, and third rank 1 wrong predictions (i.e., false-positives), respectively. Finally, we counted the number of rank 1 true-positives with scores above Th1, Th2, and Th3, (shown in Table IV). We excluded from the true-positives count the five family-level targets (category 1; the number of rank 1 correct predictions above Th1 within category 1 is shown separately in the last column of the table). The magnitude of the scores of each method vary depending on the scoring system used. For some methods a large positive score is a good result, and for the others, a large negative number is a good result. Increasing or decreasing values of Th1, Th2, and Th3 give an indication as to what is considered a better score for each method. Confidence thresholds help the user of an automated method to determine the reliability of a prediction. Table IV shows that for the CASP-3 targets, the selectivities of the methods were not high and that no method is able to predict much beyond the family-level targets. The implications of this to automated fold recognition are further discussed below.

TABLE IV. Selectivity of the Methods[†]

	Th1/ trues	Th2/ trues	Th3/ trues	Family-level trues above Thl
Karplus1	−18.9/0	−13.7/0	−12.6/0	4
Karplus2	−11.9/2	−8.1/3	−7.8/3	5
Karplus3	−29.9/2	−12.6/3	−9.2/7	4
frsvr_SDP	5.81/0	5.28/0	5.24/0	5
frsvr_SDPMA	7.77/0	5.83/0	5.65/0	4
frsvr_SDPMA2	5.88/0	5.03/1	4.70/2	5
GenTHREADER	0.76/3	0.69/5	0.58/5	5
3D-PSSM	0.54/1	0.61/1	0.61/1	2
(1D + 3D)-PSSM	0.43/1	0.68/2	0.71/3	3
BASIC	34.8/0	25.6/0	6.11/7	5
Topits	4.71/0	4.67/0	4.25/0	2
pscan	10.2/0	8.07/1	6.52/2	4

[†] In CAFASP-1, a perfect selectivity would be 21 "trues" above Th1 (as well as above Th2 and Th3) plus five family-level "trues." An illustration on how to read the table is given for GenTHREADER: Of the 28 rank-1 predictions evaluated (26 targets with known folds + 2 targets with novel folds), GenTHREADER's first and second false-positives had scores of 0.76 and 0.69, respectively; excluding the five family-level targets, three of its rank1 predictions had scores above 0.76 (and all three were correct), and six of its rank1 predictions had scores above 0.69 (of which five were correct).

DISCUSSION

At the very outset it is important to emphasize the difficulty in comparing the servers' results shown here in CAFASP-1 with those from fold recognition in CASP-3. First and foremost, CASP-3 was a blind trial, and any work carried out in retrospect is not. CAFASP-1 must be considered as an exercise in benchmarking rather than verifiable blind prediction. Second, in the CASP-3 experiment, although up to five predictions could be submitted, only the first model was actually assessed. Third, for multidomain targets results were submitted also for the individual domains according to the boundaries apparent from the observed 3D structure. Finally, in CASP-3 the evaluation considered alignment accuracy in addition to fold assignment. Without consideration of alignment accuracy it is quite possible that some of the predictions assumed to be correct in CAFASP-1 may produce very poor models, even to the extent that the correct domain match may be missed entirely. In such cases it may be assumed that the correct fold has been found mostly by chance.

Not every aspect of CASP-3 is ideal, however. Submissions in CASP-3 often included manual input based on structural or functional interpretation of the results of algorithms. For example, given that a particular target was known to bind DNA, groups were quite free to ignore any highly ranked fold which was not also known to bind DNA. In CASP-2, the best fold-recognition results were entered by a group that did not use a fold-recognition algorithm, but instead relied entirely on evolutionary inferences based on the known function of the target proteins.[12] In CAFASP-1, however, the fold assignments have been made exclusively by automatic servers, without any human interpretation of the results.

Accepting the fact that CAFASP-1 was not a blind test, entrants in CAFASP-1 were permitted to augment their template libraries with an entry from a recently deposited set of protein coordinate entry if that was required to ensure there was a correct hit in their library (although not all the participants took the opportunity to augment their libraries). The evaluation process excluded newer entries that are not included in the latest release of the SCOP database, which roughly corresponds to the structures available at the time CASP-3 predictions were filed. The object of CAFASP-1 was to evaluate the algorithms and not how recently each group had updated their fold libraries. Of course, in real applications, the ease of updating the library is an important aspect of the utility of each method. Clearly a server which is frequently updated will have a significant advantage over a server using an out-of-date template library. Furthermore, entrants have been able to further develop their methods over the 6 months since the last CASP-3 prediction deadline.

One of the conclusions from this study is that no single approach is markedly superior to the others evaluated when considered across the entire range of targets. Some methods performed better at the superfamily level, others at the domain level. As in CASP-3, all methods in CAFASP-1 performed poorly at the fold-level category; the differences between the methods in this category may not be statistically significant. However, the relatively small number of targets does not allow to draw more general conclusions.

It is important to stress that this work does *not* attempt to show that automated predictions are better 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. Assessing the performance of automated methods is of utmost importance to evaluate their applicability at genomic scales (see below). CAFASP-1 results *do* show that *automated* fold recognition methods perform better than *automated* pairwise sequence alignment, but for many of the harder cases, the scores may not be significant enough to allow a user to distinguish true- from false-positives.

One limitation with the variety of methods tested in CAFASP-1 is that no results have been included from pair potential-based threading methods (e.g., refs. 13, 14). Most of such threading methods are not available as servers, but are distributed as stand-alone software packages which must be installed on the user's own machine. In the assessment of fold-recognition results in CASP-3, three of the six groups selected to present their results made use of this type of fold recognition (see their corresponding reports in this issue). Unlike the classic potential-based threading methods, all of the methods in CAFASP-1 explicitly make use of the sequence information in one form or another. Several of the methods in CAFASP-1, including the top performers, also incorporate some structural information from the available coordinates. To what extent their superior performance stems from their use of structural information or from other factors (such as better alignment algorithms or better statistical scoring measures) remains to be determined.

One area in which the CAFASP-1 results are of particular interest is that of structural genomics. Automated approaches for fold recognition are essential if the wealth of data in genomes is to be exploited (e.g., refs. 15–17 and 18 for a recent review). One important aspect of genomic fold assignment, however, is that folds must be assigned with a high degree of confidence. Even if a method frequently ranks correct folds in top place, if the scores for these matches are not significant then the results will be of little use for genome annotation. To assess this aspect of fold assignment it is necessary to evaluate how well a method discriminates correct match scores from incorrect ones. Table IV shows that automatic fold-recognition methods are just beginning to discriminate correct from incorrect matches. Although a number of true-positives were identified above the first false-positive threshold (Th1), their scores do not necessarily lie above the methods' recommended confidence thresholds (see last column of Table II). A major conclusion from CAFASP-1 is that improvements in this aspect are required to allow a much wider applicability of automated fold-recognition methods at a genomic scale.

Beyond genome analysis, automated fold recognition servers enable the wider community ready access to the software. It is therefore essential that the accuracy of automatic methods of fold recognition are evaluated to allow users to decide which methods are most reliable. As a byproduct of CAFASP-1, we are planning to make available a "CAFASP meta-server," which will allow users to submit a sequence and automatically receive the results from the servers evaluated in CAFASP. The CASP experiment has already highlighted the value of blind trials; of course this must be extended to CAFASP. Although the results discussed here are not from a blind trial, we consider that one important aspect of this study is to explore what kind of strategy is required for comparative blind trials of automated structure prediction. Although CAFASP-1 concentrated on a limited evaluation of fold-recognition methods, we intend for CAFASP-2 to perform a more comprehensive evaluation and to include automatic assessment of the two other major categories of protein prediction, namely homology modeling and ab-initio methods. This will provide the community with valuable insights into the abilities and limitations of automated protein structure prediction.

ACKNOWLEDGMENTS

We thank J. Moult for his interest and support and for inviting us to join this Special Issue of *Proteins;* the many users of our servers and the many CASP-3 participants who encouraged us to bring about this event in very strict time constraints; Christine Orengo for her assistance in determining the correct hits for each target; the experimentalists that permitted their sequences to be used as benchmarks in CASP-3 (and CAFASP-1); and last, we thank the Internet technology advances that allowed CAFASP-1 to take place. L.A.K. is supported by GlaxoWellcome. L.A.K., R.M.M., and M.J.E.S. thank Dr. Saqi (GlaxoWellcome) for helpful discussions. A.G. and K.P. are supported by NIH grant GM60049. K.J.K. is supported by NSF grants BIR-9408579 and DBI-9808077 and DOE grant DE-FG03-95ER62112.

REFERENCES

1. Murzin AG, Brenner SE, Hubbard T, Chothia C. Scop: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
2. Fischer D, Elofsson A, Rice DW, Eisenberg D. Assessing the performance of inverted protein folding methods by means of an extensive benchmark. Proc 1st. Pacific Symposium on Biocomputing. January 1996. p 300–318. http://www.mbi.ucla.edu/people/fischer/BENCH/benchmark1.html.
3. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
4. Jones DT. Genthreader: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287:797–815.
5. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
6. Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
7. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. Cath—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
8. Fischer D, Eisenberg D. Protein fold recognition using sequence-derived predictions. Prot Sci 1996;5:947–955.
9. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
10. Fischer D. Modeling three-dimensional protein structures for amino acid sequences of the CASP3 experiment using sequence-derived predictions. Proteins Suppl 1999;3:61–65.
11. Karplus K, Barrett C, Hughey R. Hidden markov models for detecting remote protein homologies. Bioinformatics 1999;14:846–856.
12. Murzin AG, Bateman A. Distant homology recognition using structural classification of proteins. Proteins 1993;16:92–112.
13. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through folding motif. Proteins 1993;16:92–112.
14. Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casari G, Sippl MJ. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. J Mol Biol 1990;216:167–180.
15. 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.
16. Rychlewski L, Zhang B, Godzik A. Functional insights from structural predictions: analysis of the *Escherichia coli* genome. Protein Sci 1999;8:614–624.
17. Huynen M, Doerks T, Eisenhaber F, Orengo C, Sunyaev S, Yuan Y, Bork P. Homology-based fold predictions for *Mycoplasma genitalium* proteins. J Mol Biol 1998;280:323–326.
18. Fischer D, Eisenberg D. Predicting Structures for Genome Sequences. Curr Opin Struct Biol 1999;9:208–211.
19. 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. In: Istrail S, Pevzner P, Waterman M, editors. RECOMB 99- proceedings of the 3rd annual conference on computational biology. New York: Association for Computing Machinery; 1999. p 218–225.
20. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C. Sequence comparisons using multiple sequences detect twice as many remote homologues as pairwise methods. J Mol Biol 1998;284:1201–1210.
21. Elofsson A, Fischer D, Rice DW, Le Grand S, Eisenberg D. A study of combined structure-sequence profiles. Folding Design 1996;1:451–461.
22. Rost B. TOPITS: threading one-dimensional predictions into three-dimensional structures. Proc. Conf. Intelligent Systems in Mol Biol, ISMB-95; 1995. p 314–321.
23. Rychlewski L, Zhang B, Godzik A. Fold and function predictions for mycoplasma genitalium proteins. Folding Design 1998;3:229–238.