

Automatic Consensus-Based Fold Recognition Using Pcons, ProQ, and Pmodeller

Björn Wallner, Huisheng Fang, and Arne Elofsson

Stockholm Bioinformatics Center, Stockholm University, Stockholm, Sweden

ABSTRACT CASP provides a unique opportunity to compare the performance of automatic fold recognition methods with the performance of manual experts who might use these methods. Here, we show that a novel automatic fold recognition server, Pmodeller, is getting close to the performance of manual experts. Although a small group of experts still perform better, most of the experts participating in CASP5 actually performed worse even though they had full access to all automatic predictions. Pmodeller is based on Pcons (Lundström et al., *Protein Sci* 2001; 10(11):2354–2365) the first “consensus” predictor that uses predictions from many other servers. Therefore, the success of Pmodeller and other consensus servers should be seen as a tribute to the collective of all developers of fold recognition servers. Furthermore we show that the inclusion of another novel method, ProQ², to evaluate the quality of the protein models improves the predictions. *Proteins* 2003;53:534–541.

© 2003 Wiley-Liss, Inc.

Key words: fold recognition; threading; LiveBench; CASP; CAFASP; protein structure prediction

INTRODUCTION

Since the first CAFASP experiment at CASP3, it has been clear that manual fold recognition experts perform significantly better than automatic methods. Obviously, if we could completely understand what methods the manual experts use, we should be able to write computer programs that use the same tricks. Here, we describe an automated method, Pmodeller, that uses two tricks, (consensus analysis and structural evaluation) to improve automatic fold recognition. Below, we show that although Pmodeller does not perform as well as the best manual experts, it performs better than most groups as well as better than servers that do not use consensus analysis.

A common trick used by fold recognition experts is to use consensus analysis. In such an analysis, not only one prediction for each target is considered. Instead, models from different methods, with similar scores, created by different parameters or from homologous sequences, are all taken into account. In contrast, most automatic fold recognition methods return a single best prediction, ignoring any information about potential similarity with alternative predictions.

Recently, we introduced an automatic consensus method, Pcons,¹ which tries to reproduce the consensus analysis

made by manual predictors. The idea of consensus analysis is to gather predictions from a set of different methods (in this case, a set of Web servers) and then compare the obtained models with each others. The consensus analysis will give higher scores to models that have many neighbors (i.e., when many other models are similar to this one). As shown before, by the semiautomatic predictions by the CAFASP-CONSENSUS² group in CASP4 and later by Pcons and other consensus methods^{3,4} in LiveBench,^{5,6} the performance of consensus methods is significantly higher than for individual methods. However, the most significant improvement is an improved specificity, which is not analyzed in CASP⁷ but in CAFASP⁸ and LiveBench.

Another trick used by manual experts is to actually examine the structure. Although many methods do use some type of structural information, most methods do not exclude bad models, (e.g., models with large gaps, a lot of hydrophobic surfaces, low secondary structure content, or other such features). To our knowledge, only GenTHREADER⁹ evaluates the quality of the final model. In its score, GenTHREADER includes the THREADER¹⁰ score, which includes information about residue–residue contacts. However, many other methods, such as 3D-PSSM,¹¹ use structural information as a part of the alignment procedure, but using structural information does not guarantee that good models are generated. We recently developed a method, ProQ,¹² that predicts the quality of a protein model by analyzing a large set of features. ProQ performs better than other predictors at separating correct from incorrect models. By combining the output from Pcons with a homology modeling procedure and ProQ, we have developed a new predictor, Pmodeller, which first was tested at CASP5.

In addition to compare Pmodeller with other automatic servers, CASP5 provides a unique opportunity for comparisons between man and machine. Below, as well as in other articles in this issue, it can be seen that automatic method based on the idea of consensus analysis, such as Pcons, still does not perform as well as the best manual predic-

Grant sponsor: Swedish Research Council; Grant sponsor: Foundation for Strategic Research; Grant sponsor: Carl Trygger Foundation; Grant sponsor: Graduate Research School in Genomics and Bioinformatics.

*Correspondence to: Arne Elofsson, Stockholm Bioinformatics Center, Stockholm University, SE-106 91 Stockholm, Sweden. E-mail: arne@sbcsu.se

Received 11 February 2003; Accepted 8 April 2003

TABLE I. Description of the Development of Pcons 1–3

	Pcons1	Pcons2	Pcons3
Training	ANN	MLR	MLR
Comparison	LG-1.0	LG-2.0	LG-2.0
Server 1	FFAS	FFAS	Fugue 2.1
Server 2	PDBBLAST	PDBBLAST	3D SHOTGUN INBGU
Server 3	INBGU	INBGU	Orfeus
Server 4	GenTHREADER	mGenTHREADER	—
Server 5	3D-PSSM	3D-PSSM	—
Server 6	—	Sam-T99	—
Server 7	—	Fugue 1.0	—

Pcons1 used five servers, whereas Pcons2 and Pcons3 used seven and three, respectively. The other major difference between Pcons1 and the later versions is that the later version did use multiple linear regression (MLR) instead of a neural network (ANN). A modified version of the structural comparison program LGscore was also used in the later versions. For a more detailed description, see Fang et al.¹³

tors. However, the best consensus predictors perform better than most experts and better than all non-consensus-based servers.

MATERIALS AND METHODS

Pcons

The Pcons^{1,13} consensus approach differs from earlier fold recognition methods, because a set of publicly available Web servers are used to produce the input data. To efficiently combine the results of various method, all collected models are compared by using a structural superposition algorithm. Pcons then tries to predict the quality of all collected models, and if several servers predict one particular fold, Pcons will assign a high score to it. Finally, the predicted model quality and the similarity to other models are being used in the ultimate assessment and scoring of the evaluated model. From several benchmarks it is clear that Pcons performs significantly better than any of the servers that it uses,⁶ both by making more correct predictions and, more importantly, by having a higher specificity. From studying the dependence of different inputs it was obvious that the improvement was due to the inclusion of structural consensus between the different models.¹

Since the development of Pcons, at least two other consensus-based methods have been developed: 3D-SHOTGUN by Fischer³ and 3D-jury by Rychlewski et al.⁴ From the LiveBench studies, it seems as if the performance of these methods is comparable. We have also developed Pcons further; see below. The predictions from Pcons are also used as the input for two other servers: ROSETTA by David Baker and coworkers¹⁴ and Pmodeller, described here.

Development of Pcons2 and 3

Between the development of the first Pcons version in 1999 and CASP5, two new versions of Pcons were developed: Pcons2 and Pcons3. The major changes in Pcons2 are that the neural networks used in version 1 have been replaced by a simpler multiple linear regression function and that a slightly different set of servers is used; see Table I. In Pcons2 predictions from the following

seven servers are included: mGenTHREADER,¹⁰ FFAS,¹⁵ INBGU,¹⁶ 3D-PSSM,¹¹ PDBBLAST,¹⁷ Sam-T99,¹⁸ and FUGUE.¹⁹ The results from the structural comparison are also reformatted to include an average similarity to all other models. For a detailed description, see Fang et al., 2002.¹³

In LiveBench-4, we observed that several new individual servers performed significantly better than the servers used in Pcons2. Therefore, we decided to develop Pcons3 by using the best of these servers. When developing Pcons3, we did not observe any improvement using more than these three servers: INBGU-3D-SHOTGUN, Fugue-3, and Orfeus. None of these servers are used in Pcons2, and they all showed a good performance in LiveBench-4; see <http://bioinfo.pl/LiveBench/4/>. This and other test results made us believe that Pcons3 should perform significantly better than Pcons2.

All first rank models predicted by Pcons originate from a particular server. In Figure 1 it can be seen that the first ranked models originate from all participating servers both in Pcons2 and 3. The two most popular methods used by Pcons2 are INBGU and PDBBLAST, whereas Orfeus is the most frequently utilized server in Pcons3.

ProQ

Recently, we developed a neural network-based method to predict the quality of a protein model ProQ.¹² The quality of a model is defined in a similar way as in Pcons, LiveBench, CASP, and CAFASP using LGscore²⁰ and MaxSub.²¹ We have shown that ProQ performs at least as well as other structural evaluation methods for the identification of the native structure and better for the detection of correct models. This performance is maintained over several different test sets. One reason why ProQ is better at detecting correct models is most likely because it was optimized for this particular task, and thus better than methods designed to find native structures. Another reason for the improvement is that ProQ uses a combination of several structural features, which seems to improve the detection of correct models. These features include atom–atom contacts, residue–residue contacts, surface area exposure, secondary structure agreement, fatness, and other

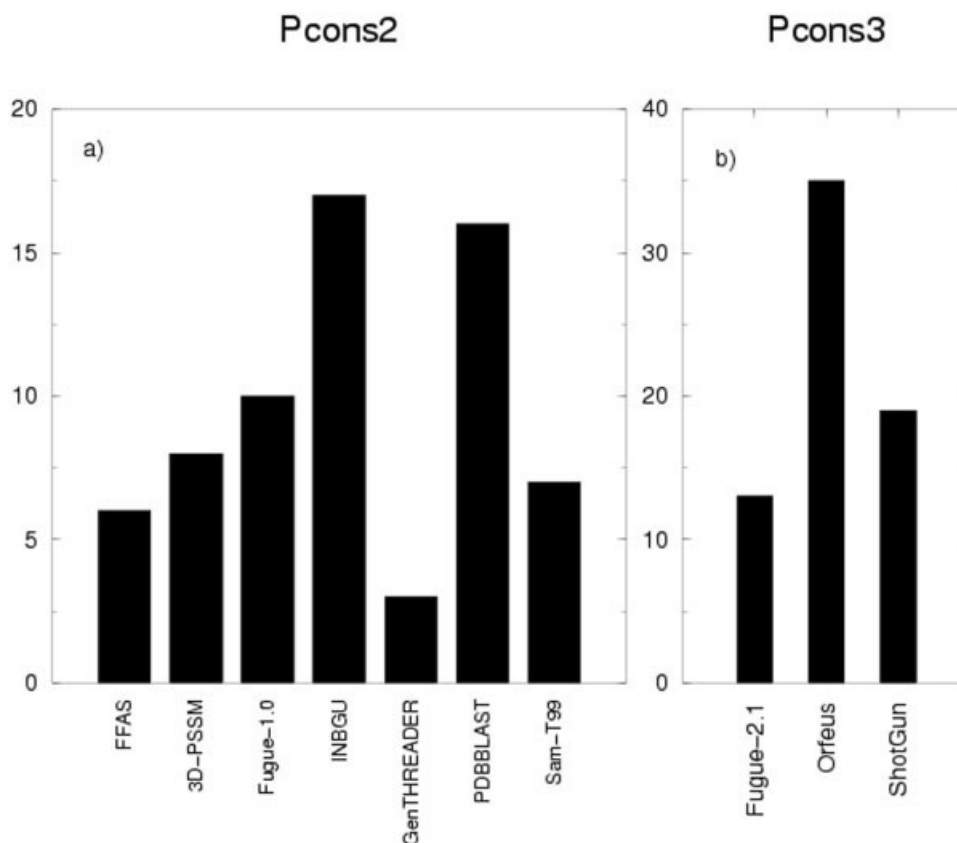


Fig. 1. The number of times the first ranked model originated from a particular method. Most models selected by Pcons2 originates from INBGU or PDBBLAST, whereas Orfeus is the most popular method in Pcons3. However, models from all methods are selected sometimes. A similar distribution is obtained by Pmodeller.

factors. By including all these factors, we were able to improve the correlation between predicted and observed quality from <0.5 for a single factor to 0.75 when all factors were combined.

Pmodeller

Pmodeller is a combination of Pcons and ProQ; see Figure 2 for a description. The alignments from the different servers are turned into all-atom models using MODELLER-6²² with predicted secondary structure information from PSIPRED²³ and then evaluated both by Pcons and ProQ. For each model, the quality is predicted by a simple linear combination of the ProQ and Pcons quality predictions:

$$Pmodeller\ score = 0.17ProQ + 0.85Pcons$$

where *ProQ* represents the prediction of the quality by ProQ and *Pcons* represents the Pcons score. Both Pcons and ProQ are trained to predict the LGscore²⁰; therefore, the scores are of comparable sizes. Thus, the *Pmodeller_score* is dominated by the Pcons prediction. Its main advantage over Pcons is that a number of high-scoring false-positive models could be eliminated, resulting in higher specificity. In CASP5, we used two different Pmod-

eller versions: Pmodeller, which is based on Pcons2, and Pmodeller 3, which uses Pcons3.

RESULTS

Below we compare the performance of Pcons and Pmodeller with each other as well as with other groups/servers. We chose to only focus on the first prediction for each target because many manual groups only submitted one, or a few, predictions for each target. Therefore, including more than one prediction would not provide a fair comparison between automatic and manual predictors.

In general, our evaluation is quite similar to the evaluation performed elsewhere in this issue. To analyze differences in performance, we used GDT_TS²⁴ scoring for the different models. For additional analysis, the AL0 and AL4 scoring were also used. We have made two assumptions in our analysis. The first was to ignore insignificant differences in performance we assume that if two models differ with <5 in GDT_TS score (i.e., $<5\%$ difference in the number of superposable residues), there is no significant difference between them. Second, we ignored models where none of the compared methods made a “correct” prediction. We found that alignments of short fragments are often due to the alignment of long helices or super secondary structure elements and that such alignments do not correlate

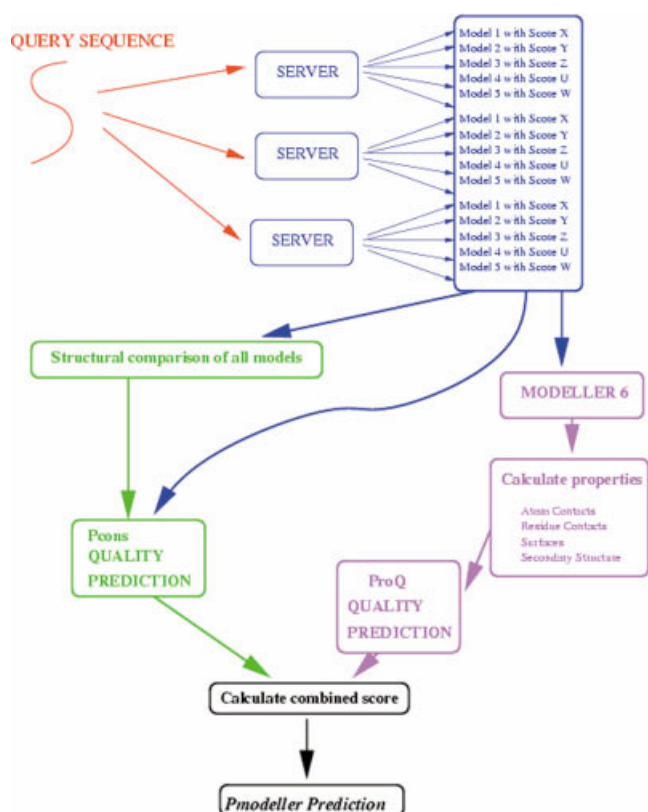


Fig. 2. The Pmodeller Scheme. A query sequence (in red) is sent to a set of fold recognition servers by the meta server.²⁶ The output from these servers (marked in blue) is transformed into $C\alpha$ -models and the scores, ranks, and alignments are saved. The structure of the models are compared with each others, and the output of this comparison is sent to Pcons,¹ which also receives information about the scores from the servers. Pcons then makes a prediction of the quality for all models. In parallel (shown in magenta) the alignments from the servers are fed into MODELLER-6²² to create all atom models. Structural features, such as frequencies of atom-atom contacts are then calculated from these models and fed into ProQ.² ProQ predicts the quality of the models. Finally, the two quality predictions are combined, and the best scoring predictions are returned to the user.

with a structural similarity of the model and target. From studies of automatic evaluation methods, we found that if <30 atoms are aligned, there is $<50\%$ chance that the template used is from the same fold, data not shown. Therefore, we ignored all targets where none of the compared methods could align >30 residues (i.e., where the GDT_TS multiplied with the length of the query is <30). We also divided the targets into Comparative modeling targets (CM) and Fold Recognition/New Fold (FR&NF) targets in the same way that the CASP evaluators did. Below, we first compare the performance of the different Pcons and Pmodeller servers and then the performance of Pmodeller compared with manual experts and other servers.

Pmodeller Versus Pcons

From Table II it can be seen that Pmodeller performed significantly better than Pcons2 for 13 targets but worse for 4. The performance is mainly improved for CM models,

TABLE II. Comparison Between the Two Versions of Pmodeller and Pcons

	<i>Pcons2</i>	<i>Pcons3</i>	<i>Pmodeller</i>	<i>Pmodeller3</i>
All				
<i>Pcons2</i>	0	8	4	6
<i>Pcons3</i>	10	0	6	2
<i>Pmodeller</i>	13	13	0	9
<i>Pmodeller3</i>	13	4	6	0
Comparative modeling				
<i>Pcons2</i>	0	6	3	4
<i>Pcons3</i>	8	0	4	1
<i>Pmodeller</i>	11	10	0	7
<i>Pmodeller3</i>	8	2	3	0
Fold recognition and new fold				
<i>Pcons2</i>	0	2	1	2
<i>Pcons3</i>	2	0	2	1
<i>Pmodeller</i>	2	3	0	2
<i>Pmodeller3</i>	5	2	3	0

The number represents the number of times the method to the left performed significantly better, as measured by an increased GDT_TS by 5%, than the method on top.

where Pmodeller made 11 better predictions. For the FR&NF targets, there is no significant increase in performance because Pmodeller did two better and one worse predictions. Because Pmodeller is based on the predictions from Pcons2, it is possible that the predictions are based on the same alignments. In slightly more than one third of all CASP predictions, the same alignment was used in Pcons2 and Pmodeller, but only in 1 (134_2) of the 17 targets with a significant difference. In this case, the homology modeling procedure made the model worse (GDT_TS score of 55 vs 65). This finding indicates that the improvement obtained by Pmodeller is not the result of the homology modeling procedure but to the ProQ evaluation procedure. The difference in performance between Pmodeller3 and Pcons3 is smaller; only in six cases (4 for Pmodeller3 and 2 for Pcons3) is the difference in GDT_TS significant. Because Pcons3 uses fewer servers, the same alignments were used more frequently than for Pmodeller/Pcons2 (73%), but in none of the six significant different predictions was the same alignment used. This finding emphasizes again that the only reason Pmodeller performs better than Pcons is due to the ProQ structural evaluation procedure. When several models have similar Pcons scores, the ProQ evaluation might tip the score so that Pmodeller chooses a better model. However, if one model is significantly better, according to Pcons, the ProQ evaluation will not affect the results at all.

The ProQ procedure used in Pmodeller seems to produce a small, but significant improvement over Pcons. This can also be seen in the LiveBench-6²⁵ and in the CAFASP evaluations presented elsewhere in this issue. There, it is also shown that the specificity for Pmodeller is slightly higher, because ~ 5 – 10% more correct predictions are made before the first incorrect prediction (see also Table V). One difference between the results in LiveBench and the results here are that in LiveBench the improvement is largest for harder targets, whereas in CASP5 the improvement was larger for easier targets. This might be due to

either a different definition of easy/hard targets, a bias from the small sampling, or a result of different evaluation methods.

Another obvious advantage of Pmodeller, besides producing better predictions, is that Pmodeller produces all-atom models, whereas Pcons only returns alignments. When Pmodeller used the same alignment as Pcons, no examples of a significantly better model was observed; therefore, we concluded that the homology-modeling procedure does not improve models. Luckily, it does not seem to disturb the models significantly either, with one exception (134_2).

Pmodeller Versus Pmodeller3

The most noticeable difference between Pmodeller and Pmodeller3 is that Pmodeller3 uses fewer, but better, servers (Table I). During the development of Pcons3, it was shown that the performance should be increased over Pcons2, whereas here and in the current LiveBench,²⁵ it is indicated that the performance is at best comparable. This finding suggests that Pcons3 might have been overtrained. In 15 of the CASP5 targets, the difference between Pmodeller and Pmodeller3 was significant. Pmodeller performed better for nine target, whereas Pmodeller3 performed better for six (see Table II). For seven CM targets, Pmodeller performed better, whereas Pmodeller3 only did three better predictions. For the hard targets, there was no significant difference (two vs three better predictions, i.e., Pmodeller performed better for easier targets, but there is no significant difference for hard targets).

In LiveBench no significant difference between Pmodeller and Pmodeller3 can be detected because their ranking depends on what evaluation method is used. Thus, it can be concluded that Pmodeller3 does not perform significantly better than Pmodeller although it uses better servers. The lack of improvement between Pmodeller and Pmodeller3 might indicate there is an advantage to use many servers in the consensus analysis. For the models where neither Pmodeller nor Pmodeller3 made a “correct” prediction, as defined above, Pmodeller3 performed better. However, we do not believe that this difference is of any practical importance because most of these predictions most likely are wrong.

Comparison With Other Predictors

To compare the performance between Pmodeller and other groups, we used a similar approach as above (i.e., considering that two predictions are of identical quality if their GDT_TS differs with <5 and that if no one predicted a model with >30 superposable residues) it is ignored. To analyze the performance of each group, we calculated an “average rank” for each group obtained by using the formula following:

$$\frac{N}{\sum \frac{1}{Rank + 1}}$$

where N is the number of targets and $Rank$ is the number of predictions that are >5 GDT_TS units better than the current model. If a group always did (one of) the best

TABLE III. Comparison Between the Best Predictors With Servers Marked in Bold as Measured by the “Average Rank”

Method	Top (%)	Better than average (%)	Worse than average (%)
Ginalski	64	35	1
GeneSilico	52	44	4
Bujnicki J	51	29	21
BioInfo.PL	47	51	3
FISCHER	40	56	4
ORNL PROSPECT	40	56	4
Sasson Iris	43	55	3
Celltech	42	47	12
BAKER ROBETTA	36	53	10
Honig	38	40	22
TOME	38	52	10
CHIMERA	38	49	13
Jose	35	53	12
Bates Paul	31	61	8
Skolnick Kolin	31	51	18
3D SHOTGUN 3DS3	35	47	18
GeneSilico.PL S_A	32	52	16
Pmodeller	32	56	12
CBRC	34	49	17
Jones	27	60	13
Pmodeller3	27	61	12

Each row includes the name of the predictor, the percentage of predictions where the model produced one of the best predictions, a model better or worse than the average prediction.

predictions, this average rank is 1, whereas if a group always did the worse prediction, the average rank is identical to the number of groups participating in CASP5. The obvious advantage with this measure compared with a simple average is that this measure is much less sensitive to one (or a few) bad prediction. We have also counted the number of times a certain predictor did one of the best predictions (i.e., no prediction was >5 GDT_TS higher), the number of predictions better or worse than the average of all predictors (see Tables III and IV). In addition, we used the AL0 and AL4 measures to perform a similar analysis; data are available at <http://www.sbc.su.se/~arne/casp5/>. The results using AL0 are virtually identical to the ones using GDT_TS, whereas the only difference seen with AL4 is that two groups using ab initio methods (Baker and Skolnick) are ranked higher.

In Table III, the performance of the best groups sorted by the average rank is shown. It can be observed that 14 manual groups and 2 other consensus servers, ROBETTA and 3D SHOTGUN 3DS3, were ranked higher than Pmodeller. This finding is in agreement with the analysis by the official assessors. The performance of the four best groups is outstanding (making >50% “one of the best” predictions), whereas the performance of the next 20 groups is comparable (making approximately one third “one of the best predictions”) (i.e., only a few manual groups performed significantly better than Pmodeller). The Pmodeller predictions can be divided into three parts: for about 30% of the targets the Pmodeller model is one of the best predictions, for 60% it is better than average, and for the

TABLE IV. Comparison Between Server Predictions With Consensus Servers Marked in Bold, Sorted by the “Average Rank”

Method	Top (%)	Better than average (%)	Worse than average (%)
BAKER ROBETTA	58	31	11
Pmodeller	51	38	11
3D SHOTGUN 3DS5	50	35	15
3D SHOTGUN 3DS3	49	35	16
Pmodeller3	49	41	11
Pcons3	45	41	15
BioInfo.PL BscB	43	42	15
BioInfo.PL ORFe	43	39	18
3DSN INBGU	39	50	11
Pcons2	36	46	18
3D SHOTGUN INBGU	38	34	28
FUGUE3	35	49	16
SAM T02 server	36	43	20
FAMS	35	47	18
INBGU	34	51	15
FORTE1	34	47	19
FAMSD	34	50	16
mGenTHREADER	32	50	18
FUGUE2	31	50	19
FFAS03	30	49	22

The top 20 servers are shown in this table. Each row includes the name of the predictor, the number of predictions where the model produced one of the best predictions, and a model better or worse than the average prediction.

final tenth it is worse than average. In comparison, the best manual group made one of the best prediction for almost two thirds of the targets, whereas the fifth highest ranked group only made six (8%) more “best predictions” than Pmodeller. Pmodeller performs approximately equally well for CM, FR, and NF targets, whereas ROBETTA and other servers performed significantly better for the harder targets in comparison with the manual groups (see the accompanying Web site).

Using the average rank of the six best servers, Pcons, Pmodeller, ROBETTA,¹⁴ and 3D-SHOTGUN,⁴ are all consensus servers (see Table IV). For ~50% of the targets, Pmodeller made one of the best automatic predictions, whereas for only 11% of the predictions it made a worse prediction than the average server. If we divide the targets into the different CASP categories, ROBETTA outperforms the predictions made by Pmodeller for FR&NF targets, and the 3D-SHOTGUN shows a slightly higher performance on harder targets and slightly lower on the easier targets; data are not shown. In comparison with other servers, the consensus servers perform better for CM than for FR&NF targets. Pmodeller made one of the best predictions for 65% of the CM targets but only for 20% of the FR or NF targets.

Specificity

It is not only important to make many correct predictions, it is also important to know when a prediction can be trusted (i.e., it is important to compare the specificity between different methods). In CASP, difference in speci-

TABLE V. Specificity for Targets Divided by Domains According to CAFASP and LiveBench-6

Method	CAFASP	LiveBench-6
3D SHOTGUN INBG	43.0	42.9
3D SHOTGUN 3DS3	42.0	43.2
3D Jury C1	41.9	45.1
INBGU	40.8	36.9
3DSN INBGU	40.5	—
Pmodeller	40.4	47.8
BioInfo.PL ORFs	39.6	42.8
Pmodeller3	39.4	49.0
BioInfo.PL BasB	39.4	—
FUGUE3	38.6	35.3
3DJury Ca	38.4	45.9
Pcons3	36.9	44.6
3DJury A1	31.0	49.0
Pcons2	30.1	45.6

Consensus servers are marked in bold. The evaluation done uses the “atoms <3” measure, which is similar to GDT_TS.

ficity cannot be studied because a score is not assigned to all predictions. However, for the automatic predictions, we can use the evaluations done in CAFASP and LiveBench to study the specificity. Here, we have not used GDT_TS scores; instead, we have used the most similar measure from LiveBench (“atoms <3”).²⁵ Furthermore, in CAFASP, the targets are not divided into the same domains as in CASP, but in general we believe that these results should be comparable.

In Table V, the specificity, as defined in LiveBench, of the best automatic methods is shown both for the CAFASP targets and for LiveBench-6. It can be seen that Pmodeller shows a higher specificity than Pcons for both sets, again indicating that the ProQ evaluation adds value to the predictions. However, the difference in specificity is not consistent between LiveBench and CAFASP. In LiveBench, the best specificity is obtained by Pmodeller, Pmodeller3, and 3D-Jury A1, whereas 3D SHOTGUN and 3D-Jury C1 show the highest specificity in the CAFASP set. Unfortunately, we have no good explanation for this; it might be due to instabilities of some servers because they were developed during the CAFASP time. For instance, we had a small bug in Pcons2 for several weeks that might have affected its performance. It should also be highlighted that these analyses are performed on a small number of targets and that the differences between individual servers are quite small. Thus, we are pleased with the specificity of Pmodeller to be among the best for both sets, independent of evaluation method.

DISCUSSION

Clearly, the largest disappointment is that some groups still could do significantly better predictions than the best fully automatic methods. However, most of these groups actually used a consensus-based approach, such as Pcons,¹ 3D-jury,⁵ or 3D-SHOTGUN.⁴ It will be interesting to learn in more details what tricks they used to improve the predictions so that these can be included into the next generation of predictors. Another disappointment was

that Pmodeller3 did not perform better than Pmodeller, although it uses better servers. At the same time, the simpler 3D-jury method performed significantly better than Pcons in CAFASP (but not in LiveBench); data are from Leszek Rychlewski. This finding indicates that when developing consensus methods, it is probably better to include a larger number of servers than just the best ones. It should be noted that the four best groups according to our evaluation all used the 3D-jury method. Therefore, it is possible that a large part of their success was due to the excellent performance of 3D-jury on the CASP5 targets.

We are happy with the improvements made by Pmodeller over Pcons, which shows that ProQ adds value to the evaluation. Even if ProQ seems to perform better than earlier methods to distinguish correct and incorrect models, the Pmodeller score is still dominated by the Pcons score. The performance of ProQ alone is far from what would be necessary to completely distinguish correct from incorrect models. We are also pleased with the development of additional consensus methods. This finding shows that the idea of consensus analysis is viable and that several different implementations perform in a similar way.

One noticeable difference between Pmodeller and the best manual groups is that the total number of residues predicted is about 5% less for Pmodeller. In addition, the two other consensus methods, 3D-SHOTGUN and ROBETTA, predicted longer models than Pmodeller. One possible explanation for this is that Pmodeller, in its current form, is restricted to use one single alignment for each query. The ability to use more than one alignment/template per target might be one important piece still missing in Pmodeller. During the development of Pmodeller, we tried to include several alignments in the homology-modeling procedure. Although the overall prediction quality increased slightly, we noticed that for a few targets, MODELLER²² did not converge. This resulted in a few high scoring models that were completely wrong with a significant loss of specificity. Therefore, we used only single alignments in Pmodeller during CASP5. In future versions of Pmodeller we hope to be able to include multiple alignments.

One other problem is related to the problem that automatic servers lack the ability to correctly divide targets into domains. Most, if not all, fold recognition servers perform local or local-global alignments and thereby return only the hit to the best domain, whereas humans are free to divide the query sequence in any way they want. An iterative procedure, where regions without significant scores are resubmitted to the servers could be useful.

Finally, both 3D-SHOTGUN and ROBETTA perform better for hard targets than Pmodeller because both ROBETTA and 3D-SHOTGUN do not only use simple alignments for hard targets. ROBETTA uses a fragment-based assembly method for all targets without significant Pcons2 hits, whereas 3D-SHOTGUN combines fragments from several predictions using the so-called cooperative algorithms of Computer Vision. Inclusion of this type of predictions seems to be able to increase the performance of consensus predictors.

CONCLUSIONS

In this study, we show that an automatic fold recognition server Pmodeller is getting closer to the performance of manual experts. We show that a small group of experts still performs better, whereas most experts actually do not perform better. Furthermore, we show that the inclusion of a novel method, ProQ, to evaluate the quality of the protein improves the predictions.

ACKNOWLEDGMENT

We thank all developers of servers; without these, the consensus approach would not have been possible. The success of Pmodeller and other consensus-based methods should really be attributed to the whole collective force of fold recognition method developers. We hope that the development of new servers will continue; we will certainly keep on developing these ourselves. We are also grateful for the support from Leszek Rychlewski for the development of the Meta-Server. AE was supported by grants from the Swedish Research Council, the Foundation for Strategic Research, and the Carl Trygger foundation. BW was supported by a grant from the Graduate Research School in Genomics and Bioinformatics.

REFERENCES

1. Lundström J, Rychlewski L, Bujnicki J, Elofsson A. Pcons: a neural network based consensus predictor that improves fold recognition. *Protein Sci* 2001;10:2354–2365.
2. Fischer D, Elofsson A, Rychlewski L, Pazos F, Valencia A, Rost B, Ortiz A, Dunbrack R. Cafasp2: the critical assessment of fully automated structure prediction methods. *Proteins* 2001;Suppl 5:171–183.
3. Fischer D. 3d-shotgun: a novel, cooperative, fold recognition meta-predictor. *Proteins*. 2003. Forthcoming.
4. Ginalska K, Elofsson A, Fischer D, Rychlewski L. 3d-jury: a simple approach to improve protein structure predictions. *Bioinformatics*. 2003. Forthcoming.
5. Bujnicki J, Elofsson A, Fischer D, Rychlewski L. Livebench-1: continuous benchmarking of protein structure prediction servers. *Protein Sci* 2001;10:352–361.
6. Bujnicki J, Elofsson A, Fischer D, Rychlewski L. Livebench-2: large-scale automated evaluation of protein structure prediction servers. *Proteins* 2001;45 Suppl 5:184–191.
7. Moulton J, Hubbard T, Bryant SH, Fidelis K, Pedersen JT. Critical assessment of methods of proteins structure predictions (CASP): round II. *Proteins* 1997;Suppl 1:2–6.
8. Fisher D, Barret C, Bryson K, Elofsson A, Godzik A, Jones D, Karplus K, Kelley L, MacCallum R, Pawowski K, Rost B, Rychlewski L, Sternberg M. Critical assessment of fully automated protein structure prediction methods. *Proteins* 1999;Suppl 3:209–217.
9. Jones D. Genthreader: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol* 1999;287: 797–815.
10. Jones D, Taylor W, Thornton J. A new approach to protein fold recognition. *Nature* 1992;358:86–89.
11. Kelley L, MacCallum R, Sternberg M. Enhanced genome annotation using structural profiles in the program 3d-pssm. *J Mol Biol* 2000;299:523–544.
12. Wallner B, Elofsson A. Can correct protein models be identified? *Protein Sci* 2003;12:1073–1086.
13. Fang H, Wallner B, Lundström J, Wöwern VC, Elofsson A. Protein structure prediction: bioinformatics approach. In: Tsigelny IF, editor. La Jolla, CA: International University Line Biotechnology Series; 2002.
14. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss C, Baker D. Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins* 2001;Suppl 5:119–126.

15. Jaroszewski L, Rychlewski L, Godzik A. Improving the quality of twilight-zone alignments. *Protein Sci* 2000;9:1487–1496.
16. Fischer D. Hybrid fold recognition: combining sequence derived properties with evolution-ary information. In *Pacific symposium on biocomputing*. Altman R, Dunker A, Hunter L, Klien T, editors. Singapore: World Scientific; 2000;5:116–127.
17. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402.
18. Karplus K, Karchin R, Barrett C, Tu S, Cline M, Diekhans M, Grate L, Casper J, Hughey R. What is the value added by human intervention in protein structure prediction? *Proteins Suppl* 2001; 5:86–91.
19. Shi J, Blundell T, Mizuguchi, K. Fugue: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol* 2001;310: 243–257.
20. Cristobal S, Zemla A, Fischer D, Rychlewski L, Elofsson A. A study of quality measures for protein threading models. *BMC Bioinformatics* 2001;2:
21. Siew N, Elofsson A, Rychlewski L, Fischer D. Maxsub: an automated measure to assess the quality of protein structure predictions. *Bioinformatics* 2000;16:776–785.
22. Sali A, Blundell T. Comparative modelling by satisfaction of spatial restraints. *J Mol Biol* 1993;234:779–815.
23. Jones D. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999;292:195–202.
24. Zemla A, Veclovas C, Moulton J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. *Proteins* 1999;Suppl 3:22–29.
25. Rychlewski L, Fischer D, Elofsson A. Livebench-6: large scale automated evaluation of protein structure prediction servers. *Proteins* 2003;Suppl 6:542–547.
26. Bujnicki J, Elofsson A, Fischer D, Rychlewski L. Structure prediction meta server. *Bioinformatics* 2001;17:750–751.