PROTEINS: Structure, Function, and Genetics Suppl 5:184–191 (2001)
DOI 10.1002/prot.10039

LiveBench-2: Large-Scale Automated Evaluation of Protein Structure Prediction Servers

Janusz M. Bujnicki,[1] Arne Elofsson,[2] Daniel Fischer,[3] and Leszek Rychlewski[4]*
[1]*Boinformatics Laboratory, International Institute of Molecular and Cell Biology (IIMCB), Warsaw, Poland*
[2]*Stockholm Bioinformatics Center, Stockholm University, Stockholm, Sweden*
[3]*Bioinformatics, Department of Computer Science, Ben Gurion University, Beer-Sheva, Israel*
[4]*Bioinformatics Laboratory, BioInfoBank Institute, Poznan, Poland*

ABSTRACT **The aim of LiveBench is to provide a continuous evaluation of structure prediction servers to inform developers and users about the current state-of-the-art structure prediction tools. LiveBench differs from other evaluation experiments because it is a large-scale and a fully automated procedure. Since LiveBench-1, which finished in April 2000, and related but independent CASP3 and CAFASP1 experiments, significant progress in the field has occurred. Some of the new developments have already been assessed at the recent CASP4 and CAFASP2 experiments (both independently of LiveBench), but others have not been observed yet because they entail developments carried out only recently. These include the availability of new servers (Pcons, FUGUE, and Coblath) and the enhancement of previously existing tools (mGenThreader, Sam-T, and 3D-PSSM), which illustrate the fast rate at which the field is advancing. Consequently, to keep in pace with the development, we present the results of the second large-scale evaluation of protein structure prediction servers. Of the 11 fold recognition servers evaluated, two servers appear to be most sensitive. One of these is 3D-PSSM, a server significantly improved after LiveBench-1. The other top performer is the new consensus server Pcons, which significantly outperformed other servers in the specificity of predictions. LiveBench-2 shows that the top performing servers are able to accurately recognize a fold for about one third of the "difficult" targets, a clear improvement over LiveBench-1 results. Given that automated structure prediction is increasingly becoming a biologists companion, the guidelines drawn from the LiveBench experiments are likely to provide users with valuable and timely information for their prediction needs. Proteins 2001;Suppl 5:184–191.**

Key words: automated protein structure prediction; benchmarking; meta server; ToolShop; LiveBench

INTRODUCTION

The prediction of the three-dimensional (3D) structure of a protein based only on the knowledge of its amino acid sequence represents the main scientific problem approached by this project. The 3D structure provides the biologists with a powerful source of information and hints about the protein function. The structure is also a major requirement for understanding the effects of mutations, for the rational protein engineering, and for development of novel functions. Experimental methods to determine the structure of the protein are costly and slow and have little chance to keep up with the growing demand for structures caused by the massive genome-sequencing efforts and the resulting flood of protein sequences. Predicting structure with computational methods represents a complementary alternative to the experimental determination. The main problem of such in silico approach is the limited quality of models produced. This problem has been addressed for more than 30 years, and only limited success has been achieved in the struggle to improve significantly the reliability and performance of the prediction procedures.

Current procedures follow several standard approaches and use various recurring components, including initial secondary structure prediction, incorporation of evolutionary information transformed into sequence profiles, or use of scoring functions (potentials) developed for threading procedures based on the propensities of amino acids to engage in contacts with each other and with the solvent. The main difference between most methods is not new fundamental ideas about the prediction process but rather, the details of implementation and coupling of the components. However, the correct way to combine all modules remains an art, which is being advanced by leading scientific groups in exhausting manual labor. The vast amount of combinations of components, which can be applied to new methods, represents the biggest obstacle in the development plans. Only a communitywide project and automated procedures provide the perspective of feasible significant progress in acceptable period of time. The LiveBench program is aimed to provide one important contribution to the solution of this problem by creating a convenient tool for the evaluation of novel automated structure prediction procedures.

*Correspondence to: Leszek Rychlewski, Bioinformatics Laboratory, BioInfoBank Institute, ul. Limanowskiego 24A, Poznan, Poland. E-mail: leszek@bioinfo.pl

Received 9 March 2001; Accepted 10 October 2001

TABLE I. Total Ranking of 11 Structure Prediction Servers Evaluated in LiveBench-1 and LiveBench-2†

Easy				Difficult			
Name	LB1 (%)	LB2 hits (%)	(of 51)	Name	LB1 (%)	LB2 hits (%)	(of 152)
CONS		89	45.3	DALI	84	66	100.5
3DPS	53	87	44.5	CONS		33	50.8
PDBB	72	86	44.0	3DPS	17	30	45.8
ST99	52	81	41.5	INBG	17	24	37.0
MGTH		81	41.3	FFAS	10	22	34.0
INBG	58	78	40.0	MGTH		21	31.5
FFAS	72	77	39.3	FUGU		20	30.5
FUGU		76	39.0	ST99	12	17	26.5
DALI	86	72	36.8	GETH	9	17	25.3
GETH	60	71	36.3	COBL		14	21.8
COBL		51	26.0	PDBB	4	9	13.8
BLAS	16	16	8.0	BLAS	2	3	4.0

†The server names are abbreviated as follows: BLAS (Blast), PDBB (PDB-Blast), FFAS (FFAS), 3DPS (3D-PSSM), GETH (GenTHREADER), MGTH (mGenTHREADER), INBG (INBGU), ST99 (Sam-T99), FUGU (FUGUE), COBL (Coblath), and CONS (Pcons). The ranking is computed by using the number of correct predictions obtained by the servers as defined for each evaluation method averaged over all four methods used (MaxSub, LGscore1, LGscore2, and Touch). The division into "easy" and "difficult" is defined by using PDB-Blast results. A PDB-Blast prediction with E-value of <0.001 assigns the target to the "easy" category. The LB1 column shows the average percentage of correct hits (number of correct hits divided by the number of ease or difficult targets respectively) obtained by the previous version of the server in the LiveBench-1 experiment. The LB2 column shows the percentages of correct hits and the numbers of correct hits (averaged using the results obtained with four evaluation methods used, i.e., equal to the total number of correct hits as defined by four methods divided by four times the number of easy or difficult targets, 51 and 152 in LiveBench-2).

MATERIALS AND METHODS

The evaluation of protein structure prediction servers based on a large set of 203 proteins is presented in this article. The prediction targets were collected, as in the first LiveBench[1] experiment, on a weekly basis from the sets of new proteins with known structure released periodically by the Protein Data Bank[2] between April 13 and December 29, 2000. The most advanced and popular servers available to the biological community were evaluated, including Blast, PDB-Blast (a PSI-Blast[3]-based service), FFAS,[4] 3D-PSSM,[5] GenTHREADER,[6] mGenTHREADER,[6] INBGU,[7] Sam-T99,[8] FUGUE,[9] and Coblath.[10] In addition to the genuine structure prediction server, a consensus server Pcons[11] was included in the evaluation. Pcons uses the output of other servers to select the most probable correct prediction.

RESULTS

Simple Sensitivity

Sensitivity is defined as the number of correct predictions divided by the number of targets. The evaluation of the performance is divided into results obtained separately for "easy" (n = 51) and "difficult" (n = 152) targets. This reflects the standards in the evaluation procedures where targets are divided into Homology Modeling and Fold Recognition categories. One common way to provide a division between both groups is according to the PDB-Blast (or PSI-Blast) results and using the E-value of the first hit as a cutoff. The simplest way to evaluate the results is to count the number of correct predictions. A correct prediction is defined specifically for each evaluation method by using a quality cutoff of the model. Four methods were used to score the quality of the models generated by the servers: MaxSub,[12] two versions of LGscore[13] (one alignment-dependent and one alignment-independent version), and Touch (L. Rychlewski, in preparation; see Appendix), a contact overlap method that took part as an experimental evaluation method in the second round of CAFASP-2 (Critical Assessment of Fully Automated Structure Prediction) (D. Fischer, in preparation). All four methods differ in their evaluation results.

Table I shows the average number of correct prediction obtained by each evaluated server. The most sensitive servers (Pcons, 3D-PSSM, and PDB-Blast) are able to recognize 86–89% of the easy targets, whereas the rest of the servers achieve recognition rates $<81\%$. The performance on the hard targets is much lower, as expected. Here, the best servers (Pcons and 3D-PSSM) succeed only in about one third of the targets, with a group of four other servers succeeding only in one fourth to one fifth of the targets. It is evident that the fold recognition servers are considerably better than PDB-Blast, which only succeeds in 9% of the hard targets.

Table I shows also the results obtained by the previous versions of the servers in the LiveBench-1 experiment. Identical criteria for the division into easy and difficult targets were used, and the results were recalculated for LiveBench-1 predictions. The values show that most of the servers improved their performance even though not all of them were upgraded (the Blast-based prediction procedure is identical). This could be attributed to an increasing number of sequences in the database, which can be used by the services (most likely for PDB-Blast and FFAS and other similar servers), a more efficient fold library updat-

TABLE IIA. Total Ranking Based on Summarized Scores Using Various Evaluation Schemes for 11 Structure Prediction Servers Evaluated in LiveBench-2[†]

Easy						Difficult					
3 of 6		5 of 10		Blast		3 of 6		5 of 10		Blast	
CONS	1	CONS	1	CONS	1	CONS	1	CONS	1	CONS	1
3DPS	2	3DPS	2	3DPS	1	3DPS	2	3DPS	2	3DPS	2
FFAS	3	FUGU	3	FFAS	3	INBG	3	FFAS	3	INBG	3
FUGU	4	FFAS	4	ST99	4	FUGU	3	INBG	4	FFAS	4
INBG	5	INBG	5	PDBB	4	ST99	4	ST99	5	FUGU	5
MGTH	6	MGTH	6	FUGU	6	FFAS	6	FUGU	6	MGTH	6
ST99	7	ST99	7	MGTH	7	MGTH	6	MGTH	7	ST99	7
GETH	8	GETH	8	INBG	8	GETH	8	PDBB	8	GETH	8
PDBB	9	PDBB	9	GETH	9	COBL	9	COBL	9	COBL	9
COBL	10	COBL	10	COBL	10	PDBB	10	GETH	10	PDBB	10
BLAS	11	BLAS	11	BLAS	11	BLAS	11	BLAS	11	BLAS	11

[†]The server names are abbreviated as follows: BLAS (Blast), PDBB (PDB-Blast), FFAS (FFAS), 3DPS (3D-PSSM), GETH (GenTHREADER), MGTH (mGenTHREADER), INBG (INBGU), ST99 (Sam-T99), FUGU (FUGUE), COBL (Coblath), and CONS (Pcons). The results are shown separately for "easy" and "difficult" targets. Three different divisions into "easy" and "difficult" are used. The "3 of 6" division assigns a target as "easy" if 3 servers form the set including: PDB-Blast, FFAS, 3D-PSSM, mGenTHREADER, INBGU, and Sam-T99 (Servers from the previous LiveBench set) generated a correct prediction. The division "5 of 10" assigns a target as easy if 5 of all individual servers (excluding Pcons) analyzed in this LiveBench session predicted the target correctly. The "Blast" division uses only PDB-Blast to divide the targets. A PDB-Blast prediction with an E-value below 0.001 assigns the target to the "easy" category. This division can be generated before the structures of the targets are known. The table shows that the ranking can be affected if a different division of targets is used. Only the Pcons and the 3D-PSSM server obtain always the two top positions.

ing regime (especially in the case of 3D-PSSM) or an improved predictions protocol. The only server that shows clearly decreasing performance is DALI, which is not a prediction program but a structure comparison service. The decrease in the number of "correct hits" could be due to the delays in the fold library updating and is probably the result of the rapid increase in the number of known protein structures.

DALI is included in the sensitivity analysis to enable an approximation of the best possible performance of a fold recognition program, which aligns targets to templates with known fold and does not modify their structure, extracting similar parts from one template to create a model. However, as seen in the results obtained on easy targets, DALI scores rather poorly, which can be related to the outdated fold library or to procedural difficulties with structure comparison methods. Thus, the performance of DALI is clearly an underestimation of the best possible fold recognition. As seen in Table I, Pcons found >20% more correct templates for the easy targets than DALI. This could be extrapolated on the difficult target range arguing for an increase of the possible hit rate to almost 80%.

Complex Sensitivity

The disadvantage of the above approach to divide "easy" and "difficult" targets is that servers such as PDB-Blast and those that correlate with PDB-Blast because of their underlying methodology, that is, sequence-based approaches, score higher in the "easy" category and lower in the "difficult" one. To address this problem, a new division based not only on the performance of PDB-Blast but also on the performance of all servers is proposed. If enough servers predict a target correctly (based on the underlying evaluation method), the target is classified as "easy." Two cutoffs were used: 3 of 6 and 5 of 10 successful servers. The evaluation can be conducted separately for easy and difficult targets, thus using all four evaluation methods for each division strategy in easy and difficult results in 8, unfortunately not identical rankings.

TABLE IIB. Number of Easy and Difficult (HARD) Targets in the Three Divisions "3 of 6," "5 of 10," and "Blast" Obtained by Using Four Different Evaluation Methods[†]

Division	Evaluation	Easy	Hard
3 of 6	Igscore1	77	126
	Igscore2	85	118
	maxsub	84	119
	touch	77	126
5 of 10	Igscore1	64	139
	Igscore2	73	130
	maxsub	69	134
	touch	63	140
Blast	Igscore1	60	143
	Igscore2	63	140
	maxsub	60	143
	touch	48	155

[†]Because the classification of the targets depends on the number of correct predictions, it is specific to the evaluation method used. Thus, there are four different classifications of hits in each division based on LGscore1, LGscore2, MaxSub, and Touch. The number of differently classified targets between all pairs of protocols is shown in Table 2C.

The rankings can be constructed based on three criteria: (a) the total quality of the models, (b) the number of correct predictions, disregarding the quality of the model (as in Table I), and (c) the number of best models produced for all targets, which multiplies the number of rankings by 3. Many other approaches are possible, and the diverse results clearly show the difficulty of comparing prediction results from different sources even if the test set is large enough.

To address this problem, a consensus evaluation can be applied. The performance of two servers is compared by using all 12 ranking schemes (4 evaluation methods multiplied with 3 evaluation criteria), and the final relation is defined as the number of ranking schemes where the first server is "better" than the second. One server is clearly better than the other if it is better based on all 12 rankings. The calculated pair-wise comparisons between servers can be used to compute a total ranking, where the total rank is defined as the number of servers that are better (plus 1; a total rank of 1 means that no other server is better). Table IIA shows the ranking obtained on the "easy" and "difficult" targets, when using three different criteria to separate the targets (Tables IIB and IIC show the differences in the classification of targets using 12 different division schemes). The variations of rankings show the important effect of the division on the ranking of servers. The results also confirm the effect of the division on the PDB-Blast performance. By using PDB-Blast E-value independent criteria, most other prediction servers show superior performance on the "easy" targets compared to PDB-Blast.

Specificity

The specificity of structure prediction servers is extremely important in large-scale automated sequence annotation experiments. Two servers with similar sensitivity but different specificity have a dramatically different utility. If the confidence levels reported by the servers would not correspond to the observed prediction reliability and would have to be disregarded, serious doubts about the application of such servers could be raised. Taking into account the current sensitivity level of about 30% for the difficult targets, all predictions (returned models) would have twice as high probability to be wrong (70%) than to be correct (30%). In contrast, an optimal procedure of confidence level assignment would give higher scores to all correct predictions than to the wrong ones. If the main goal of the large-scale analysis is to select few strong "leads," the specificity of the servers (the reliability of the reported confidence scores) is a critical feature. Tables IIIA and IIIB show the results of this analysis for all participating servers. Numbers of correct predictions with higher confidence scores than each out of the first 10 false predictions are listed. These data enable verification of how well a server separates confident, correct predictions from not confident, possibly wrong ones. To provide just one value for each server, the specificity scores are averaged over the numbers obtained for the first 10 "false positives." Tables IIIA and IIIB confirm the observation from earlier experiments that the most sensitive independent servers (3D-PSSM) are not necessarily the most specific ones. According to our analysis, mGenTHREADER outperforms other individual servers in this respect. The results illustrate also probably the most important finding of our analysis. The consensus server Pcons is much more specific than the individual servers. The difference in performance is very strong, and the high specificity of Pcons is probably the most important feature of the new service.

TABLE IIC. Number of Differently Classified Targets Between Pairs of Evaluation Protocols[†]

3 of 6-Igscore1	3 of 6-Igscore2	3 of 6-maxsub	3 of 6-touch	5 of 10-Igscore1	5 of 10-Igscore2	5 of 10-maxsub	5 of 10-touch	Blast-Igscore1	Blast-Igscore2	Blast-maxsub	Blast-touch	
	8	7	28	13	12	16	30	25	28	31	41	3 of 6-Igscore1
8		5	30	21	12	18	36	29	28	35	49	3 of 6-Igscore2
7	5		27	20	11	15	33	28	31	32	46	3 of 6-maxsub
28	30	27		25	28	24	14	43	46	39	33	3 of 6-touch
13	21	20	25		9	7	19	24	27	28	32	5 of 10-Igscore1
12	12	11	28	9		8	24	23	26	29	41	5 of 10-Igscore2
16	18	15	24	7	8		18	27	30	27	35	5 of 10-maxsub
30	36	33	14	19	24	18		39	42	33	27	5 of 10-touch
25	29	28	43	24	23	27	39		3	6	22	blast-Igscore1
28	28	31	46	27	26	30	42	3		9	25	blast-Igscore2
31	35	32	39	28	29	27	33	6	9		16	blast-maxsub
41	49	46	33	32	41	35	27	22	25	16		blast-touch

[†]Even if the number of easy and difficult targets is similar between two methods, the classification of each targets can be quite different. Table 2B shows, for example, that the number of easy hits using the 3 of 6 division is 77 for both LGscore1 and Touch. However, Table 2C shows that 28 targets are assigned differently by both methods, which demonstrates a substantial (14%) difference between the assignments.

TABLE IIIA. Specificity Computed on All Targets for the Evaluated Servers[†]

Name	Mean	False	1	2	3	4	5	6	7	8	9	10
CONS	79.8	Correct	58	72	80	82	82	83	83	85	86	87
		Raw	5.75	4.65	4.33	4.17	4.15	4.03	3.94	3.49	3.35	3.32
MGTH	53.8	Correct	44	49	51	54	54	54	55	58	58	61
		Raw	0.962	0.849	0.811	0.747	0.71	0.639	0.603	0.544	0.542	0.513
FFAS	52.1	Correct	39	52	53	53	53	53	53	55	55	55
		Raw	14.11	7.65	7.53	7.52	7.45	7.44	7.43	7.3	7.28	7.23
PDBB	47.2	Correct	38	45	46	48	49	49	49	49	49	50
		Raw	1.E-07	9.E-04	0.006	0.024	0.033	0.035	0.057	0.063	0.072	0.074
INBG	47.0	Correct	27	32	38	46	52	53	55	55	56	56
		Raw	28.4	25.8	23.1	18.9	17	16.8	16.6	16.5	16.1	16
FUGU	44.7	Correct	35	41	41	44	44	48	48	48	48	50
		Raw	8.13	5.92	5.86	5.7	5.62	5.34	5.31	5.15	5.05	4.96
3DPS	44.6	Correct	31	33	45	45	45	47	47	47	49	57
		Raw	0.027	0.049	0.209	0.230	0.233	0.267	0.282	0.305	0.341	0.534
GETH	38.6	Correct	31	34	36	37	40	40	40	40	43	45
		Raw	0.962	0.899	0.878	0.849	0.747	0.73	0.71	0.696	0.603	0.548
ST99	35.8	Correct	22	23	27	36	36	37	39	44	47	47
		Raw	44.3	34.2	27.3	21.3	18.7	17.9	16.7	13.1	12.6	12.5
BLAS	3.6	Correct	0	3	3	3	3	4	5	5	5	5
		Raw	0.11	0.15	0.15	0.16	0.18	0.19	0.2	0.22	0.25	0.27

[†]For each server listed in column name (using abbreviations as in previous tables), the numbers of correct predictions are shown in rows "Correct," which have higher raw reliability score (as reported by the server) than the reliability scores for the first 10 false predictions (columns 1–10). For each false prediction, the raw reliability score reported by the server is printed in rows "Raw." An average number of correct predictions based on the 10 numbers of correct hits is calculated and used to rank the servers. A prediction is ranked as "correct" or "wrong" only if three of the four evaluation methods agree (predictions where two methods say "wrong" and two methods say "correct" are removed) and if the model is longer than 25 residues (short prediction are removed). The table demonstrates the clear superiority of the consensus approach in the specificity of the predictions.

TABLE IIIB. Specificity Computed Separately on Easy and Difficult Targets for the Evaluated Servers[†]

Easy												Difficult											
Name	Mean	1	2	3	4	5	6	7	8	9	10	Name	Mean	1	2	3	4	5	6	7	8	9	10
CONS	45.6	44	46	46	46							CONS	36	18	35	36	36	37	37	39	40	41	41
PDBB	42.7	38	45									MGTH	18.4	11	16	16	16	19	19	21	22	22	22
3DPS	41.7	25	44	45	45	45						INBG	16.3	6	7	11	15	18	19	21	22	22	22
MGTH	39.7	36	38	38	40	41	41	41	41			FFAS	15.4	5	15	15	17	17	17	17	17	17	17
INBG	39.1	34	34	38	40	40	41	41	41	41		GETH	11.8	6	7	10	10	12	14	14	14	15	16
FFAS	38.1	38	38	38	38	38	38	38	38			3DPS	11.7	6	10	10	10	11	11	11	12	18	18
FUGU	38.1	30	35	36	40	40	40	40	40	40	40	FUGU	11.1	6	6	9	9	13	13	13	14	14	14
ST99	33.7	32	34	34	34	34						ST99	7	2	2	2	5	5	5	10	13	13	13
GETH	32.1	27	29	30	30	33	34	34	34	35	35	PDBB	3.9	1	3	4	4	4	4	4	5	5	5
BLAS	5.9	3	5	5	5	6	7	7	7	7	7	BLAS	0.4	0	0	0	0	0	0	0	1	1	2

[†]For each server listed in column Name (using abbreviations as in previous tables), the numbers of correct predictions are shown, which have higher raw reliability score (as reported by the server) than the reliability scores for the first 10 false predictions (columns 1–10). An average number of correct predictions based on the 10 numbers of correct hits is calculated and used to rank the servers. A prediction is ranked as "correct" or "wrong" only if three of the four evaluation methods agree (predictions where two methods say "wrong" and two methods say "correct" are removed) and if the model is longer than 25 residues (short prediction are removed). On easy targets, the number of 10 false predictions is not reached by most methods, resulting in empty fields. The results obtained on easy targets alone are strongly influenced by the completeness of the fold library.

Table IIIA contains also the values where false predictions were observed. This can help future users to estimate the reliability of their prediction results. More details and links to the false predictions are available on the web site (http://bioinfo.pl/LiveBench/).

CONCLUSIONS

Details of the methodology and details of the results obtained with all four individual model evaluation methods using many different ranking schemes can be viewed on the LiveBench www site (http://bioinfo.pl/LiveBench/). The detailed analysis of all results allows the following conclusions:

1. It is now clear that the completeness of the library used by the server strongly determines the overall performance, not only on the "easy" targets but also on the "difficult" targets. This can be shown by comparing the

performance of FFAS with a library of folds from the beginning of the LiveBench-2 cycle with the performance evaluated after completion of the LiveBench data collection process (results available on http://bioinfo.pl/ToolShop/).
2. The 3D-PSSM server shows superior sensitivity compared to other individual servers (excluding the consensus server, which uses the 3D-PSSM results). This reflects a major improvement of the servers compared with previous benchmarking results. This can be partially attributed to the scrupulous scheme of updating the fold library but also to the additional features added to the server (including the functional analysis performed by SAWTED[14]).
3. mGenTHREADER shows the highest specificity among other individual servers (excluding the consensus server). This finding confirms the thesis that servers that failed to obtain highest ranks based on the sensitivity in selecting correct templates can be superior if the ranking is based on the specificity.
4. Pcons, the first attempt to automate the joint use of several prediction servers, showed promising performance. It combines the sensitivity of 3D-PSSM with a very high specificity. The most important feature contributing to the improved performance of Pcons is the measure of structural similarity between top hits. This feature is not sufficiently used by most fold recognition servers. From earlier CASP experiments,[15] it has been a common practice among prediction teams to take into account several models obtained with any method and select the most common fold among the top predictions. Pcons follows this strategy in an automated manner. The server performs very well in picking the correct answer; however, it often fails to pick the best one. Thus, it receives high ratings in the number of correct hits, whereas it is sometimes not better than 3D-PSSM if the total quality of the models is used as criteria. Nevertheless, Pcons represent probably the most significant progress in the field of automated protein structure prediction. Of course, it has to be clearly noted that Pcons uses the results obtained by other servers including 3D-PSSM, and all contributing servers deserve credit for the performance of Pcons.
5. The division of targets in "easy" and "difficult" category strongly affects the conclusions about the performance of servers (data available on the LiveBench www site). By using the PDB-Blast correlated division, Sam-T99 showed limited applicability in the "difficult" target range, because most of the good predictions correlated with PDB-Blast results and were artificially moved to the "easy" category. Based on another division, which uses the total performance of the group of servers, Sam-T99 shows very good performance in the "difficult" target range. This supports the thesis that the difference between many investigated servers is not statistically significant. Only the superior sensitivity of Pcons and 3D-PSSM compared to other servers seems not to be strongly affected by the division.
6. FUGUE, a new server added to the evaluated group since LiveBench-1 obtained very good results.
7. In the CASP-4 experiment, which was held approximately at the same time as the results for the LiveBench-2 evaluation were collected, the 3D-PSSM server ranked highest. Pcons did not take part as an automated server in CASP-4. Thus, the high performance of 3D-PSSM relative to other servers evaluated in both programs can be confirmed by the LiveBench experiment. The results published by the CASP-4 evaluators differ from the results obtained by the CAFASP team. However, because of the limited set of targets in CASP and CAFASP (ca. 40) and because of the differences in the evaluation protocol as well as in the details of the classification of the targets, significant differences in ranking are expected. As discussed in the "Complex Sensitivity" section, differences in ranking are expected even on a large benchmark, if evaluation protocols differ.

The LiveBench program has several shortcomings. It is very difficult to estimate the accuracy of the structure prediction algorithms that are used by the servers because their performance is strongly affected by the completeness and the updating protocol of the used fold library. It also shows that for a pair of servers there are evaluation procedures that result in contradictory conclusions about the mutual performance differences. On the other hand, the benefits of continuous evaluation of prediction servers for the community can be found in the subtle pressure to improve the services, in the general assessment of the reliability of current prediction methods, and in the progress regarding the manual structure prediction procedures and use of servers. The current LiveBench program is also equipped with automated periodical evaluation procedures that enable continuous monitoring of the progress of the experiment and the resulting changes in the rankings.

The LiveBench program clearly indicates the progress in the field of structure prediction. Compared with the previous experiment, several servers have been updated or improved, including Sam-T99, mGenTHREADER, and 3D-PSSM, whereas others were left unchanged and provide a good reference point for the progress. Sam-T99 clearly performs better than the previous version Sam-T98 when looking at the rankings obtained. mGenTHREADER shows clear superiority over the previous GenTHREADER and advanced to the most specific server (excluding Pcons). 3D-PSSM has been updated, and with the help of the new library-updating regime, it moved to the undisputed first position in the sensitivity ranking of the individual servers (excluding Pcons). In addition, new servers have been incorporated. Especially FUGUE represents a very useful addition to the set of tools for biologists. However, the main progress is based on the idea of combining several servers under a jury system. Pcons represents the first very successful implementation of this idea. The server managed to combine the highest sensitivity with the highest specificity observed among the servers, creating

the most powerful automated protein structure prediction method.

Despite the authors' confidence in the utility of Pcons, it has to be clearly stated that several methods evaluated by LiveBench are also developed by or closely related to the directors of the LiveBench program. They include Pcons, INBGU, FFAS, and also partially PDB-Blast (or Blast in general) because of the selection of run time parameters and because of the preparation of customized databases. Thus, the evaluation of these servers cannot be regarded as independent.

DISCUSSION

The ideal design of a machinery to produce scientific results in the field of protein structure prediction would involve an automated program that provides new modules and procedures for the predictions of partial structural information. The modules would be then tested and their individual utility would be evaluated by the second program. The third program would combine the modules by using statistical procedures, neural networks, or jury systems into a superior prediction method. Some of the components of this ideal machinery have been already initiated and include communitywide evaluation programs such as LiveBench[1] as well as the Consensus Structure Prediction Method Project (L. Rychlewski, manuscript in preparation).

In the current setting, the first automated program is still missing. The modules and procedures are provided by the scientific groups in the form of servers. Programs such as CASP, which evaluates the structure prediction performance of groups of experts, offer often an inspiration for the development of such modules. LiveBench could represent the second program, being developed by using the experience from the CAFASP projects. CAFASP is aimed at the evaluation of the performance characteristics of servers provided by the community.[16] LiveBench is the extension of the CAFASP idea and provides infrastructure for the automated communication between servers and enables automated large-scale evaluation using tools tuned in the CAFASP program. A periodically updated Consensus Prediction System (Pcons) represents the current efforts to create the final third program. All those components that represent major parts of the main infrastructure of the ideal machinery were introduced to the community and have already produced very useful outcome in hints for the improvement of individual prediction modules (structure prediction servers) and in novel jury methods that outperform all individual servers. Nevertheless, they represent demanding projects that must be continued to ensure a rash progress in the field and a perspective of a rapid development of a widely applicable and accurate protein structure prediction method. After the evaluation methods and consensus procedures will be brought to a robust level, the focus will presumably shift toward the intelligent generation of novel prediction modules, which would be equivalent to the automated production of new scientific ideas by machines. A complete new era in science would begin, if this could be achieved.

ACKNOWLEDGMENTS

We thank Adam Godzik, David T. Jones, Kevin Karplus, Spencer Tu, Lawrence A. Kelley, Bob MacCallum, Mike Sternberg, Jiye Shi, Kenji Mizuguchi, Tom L. Blundell, James Cuff, and Yibing Shan for support, valuable discussions, and free access to the protein structure prediction servers. The Pcons developers thank all developers of prediction servers for the ability to create a consensus server that bases its predictive power on the performance of all individual prediction providers.

APPENDIX

The Touch structure prediction evaluation method calculates the contact map overlap between the target and the model. This approach was successfully applied earlier for structural comparison of proteins.[17] It has been shown that alignments based on contact map overlaps are a powerful alternative to other structure-based alignments. A contact is defined between two C-α atoms if the distance between both is <8.5 Å and the sequence separation is at least 5. Local secondary structure elements do not contribute to the contact map. The score of aligning a contact is S[a][b] = 1-Ddist/3.5, where Ddist is the difference in distances between atom [a] and [b] in the target and in the model. If the difference is bigger than 3.5 Å the contact is not counted. The total score is divided by the number of contacts in the target.

The main feature of Touch and, in general, of contact map overlap measures is that it is easier to compare multidomain targets and models with each other, because the method is less sensitive to the relative position of the domains. An important contribution to the scores comes also from correct supersecondary structure assignment. Models that did not show correct topology but with correct supersecondary structure pattern obtain often scores that are above the threshold. After various internal tests, a cutoff of 10% was selected. This has a large impact on the ranking of results, especially in the case of very difficult targets, where most of the methods fail to find the correct template. In this group of targets, methods that base the predictions on supersecondary structure propensities, seem to outperform conventional threading servers. Another important feature of the method is that models that include bigger fractions of the target have a much higher chance of obtaining better scores. This feature also favors ab initio servers that often provide coordinates for all C-α atoms.

Touch has been applied in the evaluation of CAFASP 2 models (published in this issue). The results of the evaluation are available at http://BioInfo.PL/touch/. The source code of Touch is also freely available at this site.

REFERENCES

1. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. LiveBench-1: continuous benchmarking of protein structure prediction servers. Protein Sci 2001;10:352–361.
2. Bernstein FC, Koetzle TF, Williams GJ, Meyer EEJ, 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.

3. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
4. 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.
5. Kelley LA, McCallum CM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:501–522.
6. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
7. Fischer D. Hybrid fold recognition: combining sequence derived properties with evolutionary information. Pac Symp Biocomput 2000;119–130.
8. Karplus K, Barrett C, Cline M, Diekhans M, Grate L, Hughey R. Predicting protein structure using only sequence information. Proteins 1999;Suppl 3:121–125.
9. Shi J, Blundell TL, Mizuguchi K. FUGUE Profile Library Search Against HOMSTRAD http://www-cryst.bioc.cam.ac.uk/≈fugue/. 2000. (GENERIC)
10. Shan Y, Wang G, Zhou HX. Fold recognition and accurate query-template alignment by a combination of PSI-BLAST and threading. Proteins 2001;42:23–37.
11. 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.
12. Siew N, Elofsson A, Rychlewski L, Fischer D. MaxSub: an automated measure to assess the quality of protein structure prediction. Bioinformatics 2000;16:776–785.
13. Cristobal S, Zemla A, Fischer D, Rychlewski L, Elofsson A. How the accuracy of a protein model can be measured? 2001. Submitted for publication.
14. MacCallum RM, Kelley LA, Sternberg MJ. SAWTED: structure assignment with text description—enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons. Bioinformatics 2000;16:125–129.
15. Moult J, Hubbard T, Fidelis K, Pedersen JT. Critical assessment of methods of protein structure prediction (CASP): round III. Proteins 1999;Suppl 3:2–6.
16. Fischer D, Barret C, Bryson K, Elofsson A, Godzik A, Jones D, Karplus KJ, Kelley LA, MacCallum RM, Pawlowski K, et al. CAFASP-1: critical assessment of fully automated structure prediction methods. Proteins Suppl 1999;3:209–217. 17 Godzik A, Skolnick J, Kolinski A. Regularities in interaction patterns of globular proteins. Protein Eng 1993;6:801–810.