PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Assessment of template based protein structure predictions in CASP9

Valerio Mariani,[1,2] Florian Kiefer,[1,2] Tobias Schmidt,[1,2] Juergen Haas,[1,2] and Torsten Schwede[1,2*]

[1] Biozentrum University of Basel, Switzerland
[2] SIB Swiss Institute of Bioinformatics, Basel, Switzerland

ABSTRACT

In the Ninth Edition of the Critical Assessment of Techniques for Protein Structure Prediction (CASP9), 61,665 models submitted by 176 groups were assessed for their accuracy in the template based modeling category. The models were evaluated numerically in comparison to their experimental control structures using two global measures (GDT and GDC), and a novel local score evaluating the correct modeling of local interactions (lDDT). Overall, the state of the art of template based modeling in CASP9 is high, with many groups performing well. Among the methods registered as prediction "servers", six independent groups are performing on average better than the rest. The submissions by "human" groups are dominated by meta-predictors, with one group performing noticeably better than the others. Most of the participating groups failed to assign realistic confidence estimates to their predictions, and only a very small fraction of the assessed methods have provided highly accurate models and realistic error estimates at the same time. Also, the accuracy of predictions for homo-oligomeric assemblies was overall poor, and only one group performed better than a naïve control predictor. Here, we present the results of our assessment of the CASP9 predictions in the category of template based modeling, documenting the state of the art and highlighting areas for future developments.

Proteins 2011; 79(Suppl 10):37–58.

Key words: protein structure; evaluation; structure prediction; CASP.

INTRODUCTION

Template based protein modeling techniques have become a widely used tool in many life science research projects—providing three-dimensional structure information when no direct experimental structures are available. However, computational structure prediction often falls short of accuracy compared to experimental structures, and, therefore, the usability of a protein model for a specific application crucially depends on its accuracy and completeness.[1] Low resolution models of individual domains can be sufficient for some applications such as domain-based functional assignment or as search models for molecular replacement phasing in X-ray crystallography.[2–6] However, many applications require atomic models of high resolution, for example, for designing and interpreting amino acid mutations or studying molecular mechanisms such as ligand binding, transport phenomena, or catalytic processes.[1,7,8] In all of these applications, it is not sufficient to know the overall backbone structure of a protein, but the correct packing and placement of side-chains is additionally required to realistically model properties such as interactions with cofactors, ligands, substrates, or receptors. At the time of modelling the "correct" structure of the target protein is unknown, and, therefore, the expected accuracy of a model has to be estimated to determine the utility of a model for a specific application.

When predicting the structure of a protein, it is often impossible to build a model at the same level of accuracy for all parts of the target, and a compromise has to be made between highest ac-

The authors state no conflict of interest.
Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Torsten Schwede, Biozentrum University of Basel, SIB, Swiss Institute of Bioinformatics, Klingelbergstrasse 50-70, 4056 Basel, Switzerland. E-mail: torsten.schwede@unibas.ch.

Received 24 June 2011; Revised 1 September 2011; Accepted 4 September 2011
Published online 14 September 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.23177

curacy and highest coverage. In many cases, functional proteins consist of several domains or form quaternary structures of several entities to perform their functions. Often, functionally important sites are located at the interface between one or more macro-molecular partners. Knowing the complete macromolecular assembly is therefore in many cases crucial to understand the functional properties of a protein.[9–11]

The community experiment on Critical Assessment of Techniques for Protein Structure Prediction (CASP) is assessing the state of the art in modeling techniques to provide an objective account on what can be achieved today, to identify new developments which excel beyond the average performance, and to pinpoint bottlenecks in order to stimulate the development of improved methods in the future. Predictors are asked to submit predictions for the three-dimensional structures of target proteins, including their oligomeric state (in the form of different chains A, B, C, etc.) and confidence values for each atom ("B-factor" representing the expected error in Å) using the CASP format.*

We have assessed the correctness of predictions submitted to the CASP9 experiment for all targets classified as "template based modeling" (TBM)[12] by comparing the models to their experimental control structures. We examined the differences in the groups' ability to produce accurate C_α models of individual domains—in the TBM category mainly reflecting the quality of the target-template alignment—using the global distance test (GDT).[13] Assessment of full atom models was done by using a combination of three scores: two global measures (GDT on C_α atoms and GDC on all atoms)[13] and a novel local distance difference test (lDDT) on all atoms, which evaluates how correct local interactions were modeled in the predicted structures. We also analyzed to which extent the different groups were able to provide realistic confidence estimates of their own predictions. Finally, the correctness of the prediction of the quaternary structure of the targets was evaluated by assessing how faithfully the models reproduce the protein–protein contacts in the native complex.

Overall, the state of the art of template based modeling in CASP9 is high, with many groups performing well. The top ranked groups can frequently improve over a simple model based on the single best available structural template, and in nearly all cases the top groups perform significantly better than a model based on the best template identified by PSI-BLAST.[14] One group was able to provide better C_α models for challenging targets compared to the other groups, which however did not result in an overall top ranking due to limited accuracy of their all atom coordinate models. Among the methods registered as prediction "servers", six groups are performing on average better than the rest. The submissions by "human" groups were again, as in previous years, dominated by meta-predictors, with one group performing better than the others. In contrast to the overall satisfactory results in modeling isolated domains, the accuracy of predictions of the correct quaternary structure for the CASP9 targets was overall very poor, and only one group performed better than a naïve control predictor. As in previous CASPs, most of the participating groups in the TBM category failed to assign realistic confidence estimates to their predictions, and only a very small fraction of the assessed methods were able to provide accurate models and realistic error estimates at the same time. In our view, this is a serious limitation of the utility of the methods participating in the CASP experiment for the life science community since the differences in accuracy observed between "easy" and "hard" targets are orders of magnitude larger than the differences in between most of the participating methods.[15]†

In the following, we present the results of our assessment of the CASP9 predictions in the category of template based modeling, documenting the state of the art, and highlighting areas with potential for future development.

RESULTS AND DISCUSSION

CASP9 Template-Based Modeling in Numbers: Targets, Predictions, Assessment Units

During CASP9, 129 protein sequences were released as prediction targets. Twelve of them were later canceled for various reasons, one structure was released, but only used for assessment of disorder prediction, leaving 116 valid targets to be evaluated. They ranged in size from 54 residues (T0538) up to 887 amino-acids (T0543). As in previous rounds of CASP,[16–21] the targets were split into assessment units (AUs) and categorized into "free modeling (FM)" and "template based modeling (TBM)" as described elsewhere in this special issue[12]; 102 of the targets had at least one assessment unit classified as TBM. In total, 121 AUs were assigned to the TBM category, three of those were considered as borderline cases and assigned to both the TBM and the FM categories and thus were evaluated by both assessors. In CASP9, 61,665 TBM predictions in total were submitted by 176 groups, 79 of which were registered as prediction servers. The CASP format allows submission of up to five alternative models; in the TBM category only predictions assigned as Model 1, i.e., 14,659 predictions, were considered for ranking the methods.

Since CASP8, prediction targets are classified by the organizers into two categories[22]: proteins which were

*See Prediction Center Website for details on the CASP format http://www.predictioncenter.org/casp9/.

†Confidence measures are an indispensable component of any predictive methodology. CASP has a separate category for model quality estimation. However, we do not consider these methods as a viable alternative in this context since many of the MQAPs in this category are CASP specific developments requiring a large number of alternative models as input. See also the manuscript on "Evaluation of Model Quality Predictions in CASP9" in this issue for more details.

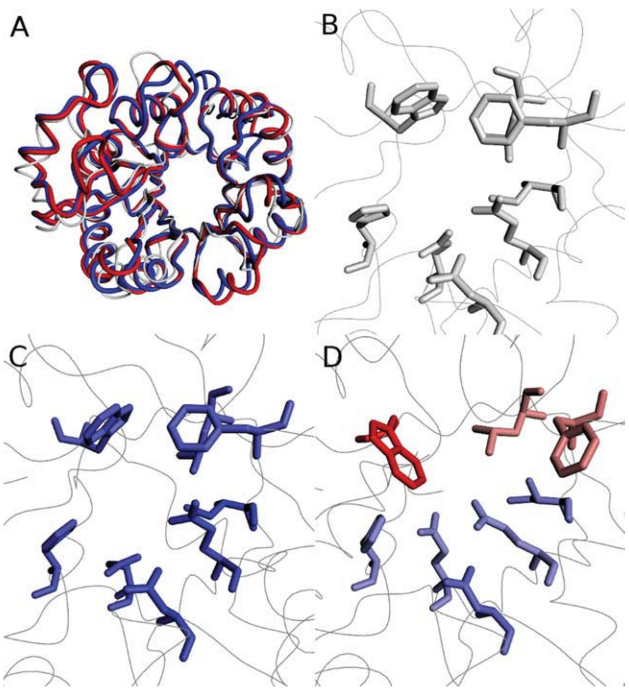


Figure 1

Comparison of the local prediction accuracy around the Glycerophosphodiester binding site in target T0570. Predictions by groups “481 Mufold-Server” and “276 RaptorX-Boost” are shown in comparison to the target structure. Panel A displays the target backbone (predictions by “481” in red and “276” in blue; target structure in gray). The backbone predictions of the two groups are very similar, as indicated by almost identical GDT-HA scores (60.9 and 60.7), while lDDT-all scores of 81.0 for “481” and 66.8 for ”276” indicate overall differences in the accuracy in predicting local interaction. Panel B shows the experimentally determined structure of the target’s binding site, and panels C and D the predictions by “481”, and “276”, respectively. Each residue is colored according to its contribution to the global lDDT-all score, using a gradient ranging from red for low scores to blue for high scores. The lDDT-all score correctly identifies group “481” as the one that predicts the structure of the local interactions more accurately. Furthermore, the residue-by-residue lDDT-all score of the binding site residues appropriately pinpoints residues Trp222, Ph158, and Leu178 as the most problematic side-chain predictions in this region.

considered as straightforward template based modeling cases were released as “server only” targets, allowing human expert groups to focus their attention on more challenging hard TBM and FM targets (“human/server” targets). Since server groups were requested to predict all targets, while human predictors were only asked to predict a subset of the targets (the “human/server” ones), the evaluation has been split into two parts: in the “server” category (S), all targets were included, but only predictions by server groups were assessed. In the “human/server” category (H/S), only human targets were considered, but predictions from all participating groups were included in the evaluation. The results of the two evaluation parts will be presented separately across most of this manuscript, but since the same assessment procedure was followed, no distinction is made in the description of the evaluation process in the following paragraphs.

Physical Plausibility of Models

As first step in the assessment, we checked if the submitted predictions actually represented physically plausible three-dimensional protein structure models. This step is essential as most numerical scores applied in the CASP assessment are based on computing numerical distances from expected atom positions but are agnostic to the physical and chemical properties of atoms. It is often possible to improve the numerical performance of distance scores such as GDT at the expense of the physical realism of a model, and the stereo-chemical quality of models submitted to CASP by some methods is deplorable. While the usage of validation tools like WhatIf,[23] ProCheck,[24] or Molprobity[25] is standard procedure in the field of experimental structural biology, stereo-chemically incorrect structure predictions are still encountered rather frequently. While earlier editions of CASP focused mainly on C_α based measures, in CASP7 all-atom measures and checks for physical plausibility were introduced,[18,21] and even stronger emphasized in CASP8.[19,20]

In the template based modeling category, all models can be expected to be physically realistic and plausible. In this round of CASP, no methods with obvious systematic compression of C_α distances were observed. We examined all predictions for steric clashes by comparing the non-bonded inter-atomic distances in the predictions to minimal reference distances observed in high resolution crystallographic protein structures. Models in which more than 5% of all residues featured impossibly short inter-atomic distances were flagged as “unrealistic”. Additionally, we used WhatIf to identify structures with serious stereochemical errors.[23] Unrealistic models were penalized in the evaluation, i.e., their Z-score was set to 0, as described in the following paragraph “Numerical Assessment”. Structures with a large number of “compressed” inter-atomic distances were mainly submitted by a small number of methods, with just eight groups accounting for more than half of all unrealistic models.

Numerical Assessment

The assessment of template based models in CASP9 was based on a numerical assessment using a combination of three different scores: GDT-HA, GDC-all, and lDDT-all. The global distance test (GDT) is a classical measure in CASP, counting the largest set of amino acid residues’ C_α positions in the prediction within certain distance cutoffs from their position in a global superposition with the experimental reference structure. Here, we apply LGA[13] to compute GDT-HA with distance cut-off values of 0.5, 1.0, 2.0, and 4.0 Å as introduced in CASP7.[18,21] GDC-all is similar in concept but takes all (non-hydrogen) atoms of the structure into account and uses ten distance thresholds (from 0.5 to 10.0 Å in steps of 0.5 Å).[20]

While GDT and GDC capture the overall displacement of atoms in a global superposition well, our visual assessment showed that these scores are not sufficient to evaluate the quality of the local atomic environment in a model (Fig. 1). We therefore introduced a new measure called Distance Difference Test (DDT). DDT rewards the fraction of correctly predicted inter-atomic distances in a model at different threshold levels, i.e., one could describe it as the "dRMSD equivalent of GDT". It can be applied to evaluating C_α positions in low accuracy predictions (DDT-CA) or to all atoms of the model (DDT-all) for TBM models. Contrary to GDT-style measures, DDT does not depend on a global superposition of the prediction and target structure and can therefore be also applied in cases where the global superposition is problematic. By using a distance threshold function, DDT can be focused on the evaluation of the local (lDDT) interactions of each atom with its neighbors (see "Material and Methods" for implementation details). For residues involved in unphysical atomic contacts (van der Waals distance violations) all distances of the residue were considered as modeled incorrectly, and since DDT rewards the fraction of correctly predicted inter-atomic distances, these residues were penalized in this measure.

Target T0570 is a glycerophosphodiester phosphodiesterase from *Parabacteroides* bacteria.[26] Figure 1 shows a comparison between two predictions ("481 MUFOLD-Server"[27] and "276-RaptorX-Boost"[28]), which have very similar GDT-HA scores (60.9 and 60.7, respectively); however, their lDDT-all scores are fairly different (81.0 and 66.8). The backbone predictions of the two groups are very similar [Figure 1(A)], as reflected in their almost identical GDT-HA scores, while significant differences are observed for their side-chain structure, as illustrated here for the glycero-phosphodiester binding site [Figure 1(B–D)]. As expected from the lDDT-all scores, the prediction by group "481 Mufold-Server" reflects the structure of the binding site in comparison to the experimental structure much better, in particular, the orientation of the tryptophan and phenylalanine side-chains. Furthermore, an analysis of the local contribution of each residue to the total lDDT-all score [highlighted with a color gradient from red for low values to blue for high values in Figure 1(C,D)] pinpoints the most problematic residues in the model of lower quality. Local atomic measures such as lDDT-all can evaluate important aspects of prediction quality which are indiscernible by the traditional global superposition based scores. lDDT complements the global assessment by a local perspective on side-chain modeling and inter-residue interactions.

For the CASP9 numerical TBM assessment, we decided to apply GDT-HA, GDC-all, and lDDT-all with equal weights, resulting overall in a higher weight for global scores (2:1) and strong emphasis (2:1) on all atom measures.

Evaluation and Ranking

As in previous CASP experiments, only "Model 1" submissions (or the one with the lowest index if no "Model 1" was available) were considered for evaluation in the TBM category. (A numerical assessment of individual models 1–5 can be found on the CASP9 web site.) For fragmented predictions, the fragment with the longest overlap with the target AU was used. The overall scoring scheme used for the assessment followed the strategy used in previous CASP editions.[18–21] For each target, GDT-HA, GDC-all and lDDT-all scores were calculated for each prediction. In order to be able to compare the prediction performance of groups across targets of varying difficulty, we calculated *Z*-scores for each of these three model accuracy measures. For each target and each score, an average value and the corresponding standard deviation was computed from all the predictions by the different groups. Models with a score more than two standard deviations worse than the average were excluded. Based on the remaining subset, average and standard deviation were re-calculated and subsequently used to compute the final *Z*-score for each prediction. All predictions flagged as unrealistic (see paragraph "Physical Plausibility of Models") were assigned *Z*-score of 0 for all three measures. As a last step, all predictions with negative *Z*-score were assigned a *Z*-score of 0. This choice made it effectively impossible to have a prediction score worse than the average for each target and was first introduced in the evaluation procedure of CASP5 to favor risk-takers and novel innovative approaches by avoiding excessive penalization for bad predictions.[29]

The computed *Z*-scores were used to rank groups according to their overall prediction performance. For each group and for each of the three evaluation scores, a median *Z*-score over all predicted targets was computed. Values for the "server" and "human/server" categories can be found in Tables I and II, respectively. Groups were finally ranked according to the average of their three median scores, and the top 25 groups were selected for a more detailed assessment. The predictions of these 25 groups were compared in a direct head-to-head analysis on common targets. In short, the raw GDT-HA, GDC-all, and lDDT-all scores of each group pairing were compared for all targets predicted by both groups using Student's *t*-test. In cases where the *Z*-score for one of the three evaluation criteria had been set to 0, the average raw score for that specific target was used. For each of the three scores, and for each group, the number of comparisons for which there was a clear win in the *t*-test (*P*-Value of less than 0.05) was summed up and the fraction of statistically significant wins was used as final ranking criterion.[21] Figure 2(A,B) shows the final ranking for the "server" and "human/server" categories. This final ranking correlates well with the preliminary ranking based on average median *Z*-scores: four of the top five groups in the "server" category are ranked in the same order, and the same is true for five of the top six groups in the "human/server" category.

In the "human/server" category, group "386 Mufold" appears to dominate the ranking with a number of wins that places it clearly ahead of its direct competitors.[30] It

Table I
Median Values and Median Z-Scores for Individual Server Predictor Groups

Group	Name	# Pred	Median GDTHA	Median GDCALL	Median LDTALL	Median Z GDTHA	Median Z GDCALL	Median Z LDDTALL	Unrealistic models	Avg Median Z
218	3D-JIGSAW_V4-0	120	45.58	53.98	70.66	0	0.07	0	4	0.02
117	3D-JIGSAW_V4-5	118	44.65	54.6	70.28	0	0.04	0	5	0.01
207	Atome2_CBS	119	46.53	55.59	72.55	0.2	0.19	0.21	1	0.2
321	BAKER-ROSETTASERVER	119	46.95	57.79	78.67	0.34	0.48	0.88	0	0.57
229	BHAGEERATH	121	9.03	9.04	51.85	0	0	0	0	0
102	Bilab-ENABLE	120	46.98	56.3	71.41	0.19	0.35	0.05	4	0.2
47	BioSerf	121	44.06	54.01	73.98	0.03	0.11	0.36	0	0.17
457	chunk-TASSER	121	47.62	57.41	58.55	0	0	0	76	0
307	chuo-fams	121	45.71	55.64	74.04	0.02	0.25	0.38	0	0.22
213	circle	115	49.24	58.14	74.64	0.2	0.35	0.37	0	0.31
142	CLEF-Server	121	46.72	54.74	75	0.17	0.23	0.4	0	0.27
264	ConStruct	90	11.8	11.29	13.96	0	0	0	68	0
214	Distill	121	43.79	54.48	56.74	0	0	0	79	0
302	FALCON-SWIFT	121	46.72	54.25	74.78	0.17	0.23	0.4	0	0.27
127	FAMSD	121	48.01	56.44	76.33	0.27	0.41	0.57	0	0.41
471	FFAS03	120	43.59	53.84	72.31	0.11	0.09	0.15	0	0.12
420	FFAS03a	121	44.83	54.83	72.18	0.08	0.08	0.15	0	0.11
476	FFAS03n	121	46.3	55.7	72.92	0.14	0.15	0.28	0	0.19
171	FFAS03ss	119	44.97	53.01	72.4	0.08	0.08	0.19	0	0.12
396	FUGUE_KM	115	40.28	29.68	35.93	0	0	0	0	0
165	GSmetaserver	117	46.71	55.21	71.59	0.07	0.17	0.19	0	0.14
236	gws	119	49.25	60.45	77.43	0.65	0.66	0.84	0	0.72
453	HHpredA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
449	HHpredB	121	53.05	61.09	77.32	0.93	0.76	0.74	0	0.81
346	HHpredC	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
275	IntFOLD-TS	121	48.78	56.82	75.58	0.29	0.34	0.48	0	0.37
63	Jiang_Assembly	121	45.96	57.37	72.35	0.13	0.3	0.1	6	0.18
366	Jiang_THREADER	121	45.86	57.37	72.6	0.17	0.3	0.11	2	0.19
355	LenServer	103	14.76	15.33	61.52	0	0	0	0	0
314	LMUserver	109	44.97	55.46	73.63	0	0	0.23	0	0.08
26	LOOPP_Austin	115	46.73	55.17	73.84	0	0.04	0.17	1	0.07
74	m4t_2009	58	50.65	60.01	75.94	0	0	0	0	0
304	Ma-OPUS-server	121	37.24	44.49	68.54	0	0	0	0	0
435	MidwayFoldingServer	117	43.55	47.83	69.35	0	0	0.01	6	0
55	MUFOLD-MD	121	40.44	51.82	74.58	0	0.07	0.62	4	0.23
481	MUFOLD-Server	121	49.25	57.69	75.89	0.42	0.42	0.62	1	0.49
2	MULTICOM-CLUSTER	121	49.75	58.69	74.22	0.56	0.61	0.4	0	0.53
80	MULTICOM-CONSTRUCT	121	49.78	58.58	74.26	0.5	0.51	0.3	0	0.44
215	MULTICOM-NOVEL	121	50.18	57.37	74.56	0.51	0.48	0.42	0	0.47
119	MULTICOM-REFINE	121	49.93	59.94	74.94	0.57	0.54	0.39	0	0.5
129	MUSICS_server	115	36.27	46.76	52.61	0	0	0	90	0
248	MUSICS-2S	93	35	40.54	63.78	0	0	0	18	0
409	MUSTER	121	44.63	53.48	74.21	0	0.13	0.37	0	0.17
436	panther	108	46.45	53.77	71.99	0.01	0.05	0.02	1	0.03
273	Pcomb	118	50.17	57.98	72.51	0.22	0.32	0.19	6	0.24
319	Pcons	118	49.83	58.28	72.01	0.18	0.32	0.05	3	0.19
208	PconsD	121	46.77	57.19	74.86	0.4	0.41	0.39	2	0.4
56	PconsM	118	49.96	59.18	72.22	0.32	0.4	0.08	4	0.26
173	PconsR	117	45.42	55.14	75.23	0	0.09	0.43	1	0.17
174	Phyre2	121	49.45	57.69	58.17	0	0	0	72	0
14	PLATO	94	11.03	10.8	53.8	0	0	0	1	0
291	prdos2	119	46.36	55.42	72.97	0	0.11	0.22	29	0.11
345	PRECORS	114	44.52	54.79	68.99	0	0	0.06	24	0.02
28	ProfileCRF	121	47.01	55.48	74.86	0.14	0.21	0.39	0	0.25
296	ProQ	111	43.43	51.88	65.97	0	0	0	9	0
1	ProQ2	118	47.49	56.35	72.21	0.18	0.32	0.21	4	0.24
253	pro-sp3-TASSER	121	45.45	57.01	55.97	0	0	0	82	0
245	PROTAGORAS	109	41.33	49.65	70.98	0	0	0.11	0	0.04
328	Pushchino	100	32.87	24.62	35.82	0	0	0	0	0
380	QUARK	121	53.47	62.12	77.32	0.75	0.68	0.75	1	0.73
286	RaptorX	121	50.99	60.66	68.74	0	0	0	55	0
276	RaptorX-Boost	121	50	60.11	62.62	0	0	0	77	0
77	RaptorX-MSA	121	52.5	60.91	72.45	0.35	0.47	0.15	35	0.33

(Continued)

Table I
(Continued)

Group	Name	# Pred	Median GDTHA	Median GDCALL	Median LDTALL	Median Z GDTHA	Median Z GDCALL	Median Z LDDTALL	Unrealistic models	Avg Median Z
350	RBO-PROTEUS	118	16.21	14.98	57.65	0	0	0	0	0
250	rehtnap	105	26.42	33.37	59.07	0	0	0	7	0
285	SAM-T02-server	112	44.73	31.93	36.52	0	0	0	0	0
244	SAM-T06-server	113	40.71	50.18	70.6	0	0	0	8	0
103	SAM-T08-server	114	46.84	56.57	72.34	0	0.22	0	17	0.07
444	schenk-torda-server	26	13.3	10.67	38.5	0	0	0	11	0
452	Seok-server	121	51.23	60.3	75.25	0.64	0.65	0.54	1	0.61
257	STAT-PROTEUS	106	12.96	10.6	44.55	0	0	0	3	0
18	Wolfson-serv	121	44.68	54.15	73.53	0	0.02	0.18	0	0.06
289	Yang_kdd	121	13.33	13.53	52.79	0	0	0	9	0
228	YASARA	72	52.76	61.79	80.43	0	0	0	0	0
428	Zhang-Server	121	53.94	62.3	77.9	0.73	0.72	0.79	2	0.75
166	ZHOU-SPARKS-X	121	47.48	57.39	73.98	0.1	0.18	0.26	22	0.18

Data shown refers to all targets.

is worth noting that according to the method descriptions (see Prediction Center website) the "human" predictor category is clearly dominated by meta-servers methods, i.e., prediction methods that use models submitted by the server predictors and either rank/resubmit them, or use them as input for further processing. It also appears that all the top performing methods are fully automated, despite being registered as "human" predictors.

In the "server" category, a group of six methods ("449 HHpredB",[31] "428 Zhang-Server",[32] "380 QUARK",[32] "452 Seok-server", "321 BAKER-ROSETTASERVER", "236 gws") clearly distinguishes itself from the rest in terms of performance, and "449 HHpredB" has a number of wins which is slightly above the others. However, compared to the "human/server" category, the groups are much closer and the top method is less dominant. One confounding factor in the final head to head comparison was that several methods with different group numbers showed nearly identical performance, which in retrospect turned out as multiple prediction services managed by the same research groups. Obviously, these services used modeling techniques with only subtle differences, and often ended up generating almost identical models. Since the CASP procedure is performed anonymously, it is not possible to recognize this situation during the assessment. We have therefore retroactively decided to exclude "453 HHpredA" and "346 HHpredC" from the assessment, because these predictions were identical to "449 HHpredB" developed by the same group.

It is worth noting that the best performing predictors, despite their overall relatively similar performance, have not converged to the same methods, but represent different unique developments with distinct strengths and weaknesses, as reflected in a diverse relative contribution of the three evaluation scores to the final performance. While the best groups predict both the overall structure and local interactions equally well, others seem to perform better at the local scale (e.g., "321 BAKER-ROSETTASERVER" or "127 FAMSD"). For other prediction groups, the opposite is true, with a relatively weak prediction accuracy of local interactions (e.g., "452 Seok-server", "2 MULTICOM-CLUSTER","119 MULTICOM-REFINE", and "215 MULTICOM_NOVEL"). This high variability suggests that the top groups use a range of different methods and approaches, actively developing the field in different directions and not simply adjusting to the most successful strategy of the previous CASP installment.

We compared the prediction performance of server and human groups on common targets (target of the "human/server" category) in order to ascertain if the longer prediction period and the availability of the models built by servers gave human predictors a clear advantage. The results were somewhat surprising as in no case the best overall human number 1 prediction was more than ten GDT-HA units more accurate than the server one. In a few cases, the server prediction even turned out to be the more accurate. It appears that servers are contributing significant value to the structure prediction problem, and the role of "human" predictions in CASP9 seems to be restricted to selection and small modifications of server models.

Improvement over PSI-BLAST Templates

At the time of modeling, the only available information about the target is its amino acid sequence. One widely used tool for detecting templates by sequence homology, which had a major impact in improving the accuracy of homology modeling in the past, is PSI-BLAST.[14] Today, several more sophisticated methods using profile HMM comparisons are commonly used.[33–36] We were interested in analyzing to which extent the use of these more sophisticated methods had an effect on template based modeling in CASP9. For this

Table II
Median Values and Median Z-Scores for Individual All Predictor Groups for Targets in the “Human/Server” Category

Group	Name	# Pred	Median GDTHA	Median GDCALL	Median LDDTALL	Median Z GDTHA	Median Z GDCALL	Median Z LDDTALL	Unrealistic models	Avg Median Z
218	3D-JIGSAW_V4-0	54	45.58	53.98	70.66	0	0	0	4	0
117	3D-JIGSAW_V4-5	52	44.65	54.6	70.28	0	0	0	5	0
300	4_BODY_POTENTIALS	51	39.9	48.16	71.63	0.32	0.28	0.55	6	0.38
311	ALAdeGAP	54	21.01	24.69	58.95	0	0	0	0	0
324	AOBA	51	46.23	54.5	76.29	0.49	0.58	0.91	0	0.66
207	Atome2_CBS	53	46.53	55.59	72.55	0	0	0	1	0
153	AuroraMBSI	33	23.36	24.11	63.88	0	0	0	0	0
172	BAKER	54	43.32	53.41	74.78	0.42	0.54	0.73	2	0.56
321	BAKER-ROSETTASERVER	55	46.95	57.79	78.67	0.59	0.59	0.94	0	0.71
424	Bates_BMM	47	35.62	42.19	71.26	0	0.08	0.36	0	0.15
229	BHAGEERATH	55	9.03	9.04	51.85	0	0	0	0	0
450	BHAGEERATH_SCFBIO	55	34.27	41.79	67.8	0	0	0	0	0
423	Bilab	54	35.97	44.59	65.21	0	0.07	0	11	0.02
102	Bilab-ENABLE	54	46.98	56.3	71.41	0.15	0.23	0	4	0.13
458	Bilab-solo	54	44.1	54	71.5	0.44	0.42	0.33	1	0.4
192	BIO_ICM	51	46.26	54.69	74.56	0.57	0.64	0.63	0	0.61
47	BioSerf	55	44.06	54.01	73.98	0	0	0.15	0	0.05
299	bujnicki-kolinski	53	46.7	56.91	73.97	0.62	0.6	0.64	0	0.62
399	Chicken_George	55	48.37	56.88	70.37	0.67	0.72	0.38	2	0.59
457	chunk-TASSER	55	47.62	57.41	58.55	0	0	0	76	0
307	chuo-fams	55	45.71	55.64	74.04	0.05	0.14	0.29	0	0.16
213	circle	49	49.24	58.14	74.64	0.1	0.13	0.2	0	0.14
220	CLEF	51	22.95	24.17	63.02	0	0	0	0	0
142	CLEF-Server	55	46.72	54.74	75	0.08	0.1	0.31	0	0.17
60	CNIO	26	35.19	43.32	68.04	0	0	0	2	0
35	CNIO-Firestar	45	42.13	52.17	71.73	0	0.08	0.23	1	0.1
264	ConStruct	42	11.8	11.29	13.96	0	0	0	68	0
301	cpu_hsfang	48	38.75	45.9	53.68	0	0	0	3	0
333	DELCLAB	51	12.16	12.17	50.94	0	0	0	0	0
214	Distill	55	43.79	54.48	56.74	0	0	0	79	0
186	Distill_human	55	36.68	39.44	42.55	0	0	0	41	0
20	dokhlab	27	19.14	19.19	59.12	0	0	0	0	0
470	elofsson	55	47.75	52.37	74.19	0.85	0.78	0.79	3	0.81
302	FALCON-SWIFT	55	46.72	54.25	74.78	0.08	0.1	0.31	0	0.17
37	fams-ace3	55	46.09	54.26	75.3	0.64	0.72	0.83	0	0.73
127	FAMSD	55	48.01	56.44	76.33	0.3	0.39	0.55	0	0.41
429	fams-multi	55	44.79	52.02	74.54	0.69	0.71	0.8	0	0.73
113	FAMSSEC	55	46.27	54.75	75.15	0.75	0.77	0.78	0	0.77
484	FEIG	50	37.05	47.44	72.56	0	0.24	0.54	0	0.26
471	FFAS03	54	43.59	53.84	72.31	0	0	0	0	0
420	FFAS03a	55	44.83	54.83	72.18	0	0	0	0	0
476	FFAS03n	55	46.3	55.7	72.92	0	0	0.01	0	0
171	FFAS03ss	53	44.97	53.01	72.4	0	0	0	0	0
83	FLOUDAS	37	29.96	33.85	63.95	0	0	0	2	0
170	FOLDIT	25	39.85	46.53	72.42	0	0	0	3	0
396	FUGUE_KM	50	40.28	29.68	35.93	0	0	0	0	0
147	GeneSilico	55	43.87	53	73.44	0.6	0.53	0.64	2	0.59
165	GSmetaserver	52	46.71	55.21	71.59	0.18	0.19	0.14	0	0.17
236	gws	55	49.25	60.45	77.43	0.39	0.53	0.76	0	0.56
42	Handl-Lovell	48	16.88	12.2	45.41	0	0	0	0	0
453	HHpredA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
449	HHpredB	55	53.05	61.09	77.32	0.73	0.68	0.64	0	0.68
346	HHpredC	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
275	IntFOLD-TS	55	48.78	56.82	75.58	0.22	0.23	0.36	0	0.27
63	Jiang_Assembly	55	45.96	57.37	72.35	0.08	0.17	0	6	0.08
366	Jiang_THREADER	55	45.86	57.37	72.6	0.16	0.36	0.01	2	0.18
104	Jones-UCL	55	49.21	56.02	76.11	0.85	0.8	0.89	1	0.85
461	jscslb	44	41.17	45.53	67.89	0	0	0.06	0	0.02
408	keasar	55	42.9	51.92	71.88	0.08	0.15	0.46	8	0.23
295	KnowMIN	55	41.77	50.1	76.78	0.67	0.6	0.88	0	0.72
382	Kurcinski-Kihara	54	42.29	52.84	73.73	0.13	0.21	0.64	1	0.33
114	LEE	55	45.62	52.48	73.89	0.73	0.68	0.8	0	0.74
361	LEEcon	54	48.31	58.37	76.67	0.76	0.83	0.95	0	0.85

(Continued)

Table II
(Continued)

Group	Name	# Pred	Median GDTHA	Median GDCALL	Median LDDTALL	Median Z GDTHA	Median Z GDCALL	Median Z LDDTALL	Unrealistic models	Avg Median Z
84	LenGroup	55	12.5	12.81	53.16	0	0	0	3	0
355	LenServer	46	14.76	15.33	61.52	0	0	0	0	0
149	LMU	47	43.79	52.87	71.04	0.11	0.09	0.31	0	0.17
314	LMUserver	48	44.97	55.46	73.63	0	0	0	0	0
26	LOOPP_Austin	50	46.73	55.17	73.84	0	0	0	1	0
316	Lovell_group	48	17.98	13.19	45.45	0	0	0	0	0
400	LTB	55	41.93	50.42	72.24	0.64	0.64	0.71	1	0.66
304	Ma-OPUS-server	55	37.24	44.49	68.54	0	0	0	0	0
94	McGuffin	55	47.75	56.06	73.6	0.79	0.74	0.55	10	0.7
200	MeilerLab	46	13.09	12.69	57.38	0	0	0	0	0
477	MidwayFoldingHuman	53	45.71	46.76	53.37	0	0	0	24	0
435	MidwayFoldingServer	55	43.55	47.83	69.35	0	0	0	6	0
386	Mufold	55	48.28	58.61	77.71	0.86	0.86	1.01	0	0.91
55	MUFOLD-MD	55	40.44	51.82	74.58	0	0	0.47	4	0.16
481	MUFOLD-Server	55	49.25	57.69	75.89	0.22	0.28	0.44	1	0.32
490	MULTICOM	55	47.34	54.95	74.23	0.82	0.82	0.71	0	0.78
2	MULTICOM-CLUSTER	55	49.75	58.69	74.22	0.41	0.54	0.27	0	0.4
80	MULTICOM-CONSTRUCT	55	49.78	58.58	74.26	0.35	0.38	0.14	0	0.29
215	MULTICOM-NOVEL	55	50.18	57.37	74.56	0.36	0.38	0.29	0	0.34
119	MULTICOM-REFINE	55	49.93	59.94	74.94	0.36	0.44	0.26	0	0.35
199	MUSICS	41	39.75	49.26	68.56	0	0	0.36	10	0.12
129	MUSICS_server	53	36.27	46.76	52.61	0	0	0	90	0
61	MUSICS-2	46	42.59	50.75	73.18	0.01	0.09	0.65	6	0.25
248	MUSICS-2S	41	35	40.54	63.78	0	0	0	18	0
409	MUSTER	55	44.63	53.48	74.21	0	0	0.28	0	0.09
436	panther	47	46.45	53.77	71.99	0	0	0	1	0
273	Pcomb	52	50.17	57.98	72.51	0.27	0.26	0.1	6	0.21
319	Pcons	52	49.83	58.28	72.01	0.29	0.32	0	3	0.2
208	PconsD	55	46.77	57.19	74.86	0.3	0.34	0.16	2	0.27
56	PconsM	52	49.96	59.18	72.22	0.27	0.29	0	4	0.19
173	PconsR	51	45.42	55.14	75.23	0	0.11	0.3	1	0.14
174	Phyre2	55	49.45	57.69	58.17	0	0	0	72	0
14	PLATO	36	11.03	10.8	53.8	0	0	0	1	0
45	POEM	53	26.44	32.8	63.95	0	0	0	0	0
291	prdos2	53	46.36	55.42	72.97	0	0	0	29	0
345	PRECORS	49	44.52	54.79	68.99	0	0	0	24	0
65	prmls	55	46.8	57.75	76.08	0.71	0.76	0.84	1	0.77
28	ProfileCRF	55	47.01	55.48	74.86	0	0.13	0.22	0	0.12
296	ProQ	48	43.43	51.88	65.97	0	0	0	9	0
1	ProQ2	52	47.49	56.35	72.21	0.02	0.16	0	4	0.06
253	pro-sp3-TASSER	55	45.45	57.01	55.97	0	0	0	82	0
245	PROTAGORAS	46	41.33	49.65	70.98	0	0	0	0	0
328	Pushchino	48	32.87	24.62	35.82	0	0	0	0	0
380	QUARK	55	53.47	62.12	77.32	0.67	0.69	0.69	1	0.69
286	RaptorX	55	50.99	60.66	68.74	0	0	0	55	0
276	RaptorX-Boost	55	50	60.11	62.62	0	0	0	77	0
77	RaptorX-MSA	55	52.5	60.91	72.45	0.33	0.41	0.07	35	0.27
350	RBO-PROTEUS	53	16.21	14.98	57.65	0	0	0	0	0
33	RecombineIt	55	48	56.42	71.24	0.74	0.77	0.55	2	0.69
250	rehtnap	45	26.42	33.37	59.07	0	0	0	7	0
285	SAM-T02-server	46	44.73	31.93	36.52	0	0	0	0	0
244	SAM-T06-server	47	40.71	50.18	70.6	0	0	0	8	0
103	SAM-T08-server	50	46.84	56.57	72.34	0	0.03	0	17	0.01
353	SAMUDRALA	53	46.81	54.21	73.27	0.62	0.54	0.62	1	0.6
360	sbtJ	55	41.38	49.85	69.49	0.39	0.42	0.4	0	0.4
182	Scheraga	31	12.91	13.5	51.7	0	0	0	15	0
242	Seok	54	47.45	55.33	70.05	0.59	0.61	0.43	0	0.54
16	Seok-meta	54	48.86	56.9	76.11	0.88	0.9	0.89	0	0.89
452	Seok-server	55	51.23	60.3	75.25	0.5	0.49	0.35	1	0.45
278	sessions	55	19.42	24.09	58.1	0	0	0	0	0
395	Shell	55	40.41	48.48	72.29	0.01	0.28	0.45	0	0.25
365	SHORTLE	52	44.94	55.32	72.13	0.65	0.6	0.41	3	0.56
391	SMEG-CCP	48	31.64	38.62	65.98	0	0	0	0	0

(Continued)

Table II
(Continued)

Group	Name	# Pred	Median GDTHA	Median GDCALL	Median LDDTALL	Median Z GDTHA	Median Z GDCALL	Median Z LDDTALL	Unrealistic models	Avg Median Z
88	Splicer	55	44.72	53.76	75.07	0.62	0.68	0.77	0	0.69
257	STAT-PROTEUS	45	12.96	10.6	44.55	0	0	0	3	0
110	Sternberg	55	46.25	55.11	59.98	0	0	0	20	0
403	StruPPi	55	33.04	39.09	65.8	0	0	0	0	0
336	SUN@tsinghua	30	10.75	9.93	38.32	0	0	0	24	0
297	SWA_TEST	54	42.57	51.83	68.62	0.33	0.27	0	11	0.2
447	SWIFT-Human	52	23.12	25.13	63.23	0	0	0	0	0
402	TASSER	55	46.23	52.69	53.05	0	0	0	31	0
282	Taylor	22	16.54	18.06	55.49	0	0	0	0	0
240	TMD3D	51	43.33	52.24	73.35	0.6	0.66	0.85	3	0.7
407	United3D	54	49.4	58.42	71.75	0.72	0.75	0.43	9	0.64
206	WAC_Labs	55	17.15	24.29	56.34	0	0	0	0	0
18	Wolfson-serv	55	44.68	54.15	73.53	0	0	0.02	0	0.01
373	Wolynes	31	14.17	10.69	43.58	0	0	0	0	0
289	Yang_kdd	55	13.33	13.53	52.79	0	0	0	9	0
228	YASARA	24	52.76	61.79	80.43	0	0	0	0	0
29	Yuan-Chen-Kihara	52	14.58	16.48	47.93	0	0	0	8	0
96	Zhang	55	49.26	54.13	75.35	0.83	0.78	0.88	1	0.83
418	Zhang_Ab_Initio	55	46.89	52.8	74.44	0.68	0.61	0.7	8	0.67
428	Zhang-Server	55	53.94	62.3	77.9	0.7	0.69	0.8	2	0.73
419	ZHOU-SPARKS-M	53	46.12	54.27	74.08	0.48	0.48	0.54	10	0.5
166	ZHOU-SPARKS-X	55	47.48	57.39	73.98	0	0	0.05	22	0.02

analysis, we used PSI-BLAST[14] to generate a sequence profile for each target sequence at the end of the prediction period and searched the PDB for suitable templates. A raw backbone model was then built by copying the backbone atoms of the corresponding template residues according to the BLAST alignment (details can be found in the "Material and Methods" section). The GDT-HA score for the resulting naïve models was used as reference to compare to the CASP9 predictions. Figure 3(A,B) show the results for the "server" and the "human/server" categories, respectively. The continuous line indicates for each target the difference in GDT-HA points of the best prediction over the corresponding PSI-BLAST template, which is represented by the 0-level baseline. (See Supporting Information Table SII for a list of targets and numerical values). In every case, there is at least one CASP9 prediction with a better GDT-HA than the naïve pseudo-model. The improvements in the "server" category range from mere 2.30 points for target T0615 to 71.26 for target T0586. This target is also the one with the largest improvement in the "human/server" category. The figures also show the improvement achieved for each target by the best individual group in red marks ("449 HHpredB" for the "server" category and "386 Mufold" for the "human" category). Also here the results are positive, with an improvement over the naïve pseudo-model for 89% of the targets by "449 HHPredB" and for 96% by "386 Mufold". In most cases, the improvement of these groups is within 10 GDT-HA points of the best overall prediction. The remarkable improvement obtained for some of the targets suggests that the predictors are often able to find a structural template that is significantly closer to the target structure than the one identified by PSI-BLAST, or to identify templates where PSI-BLAST fails. However, one should also keep in mind that the coverage of the target has a big influence on the GDT-HA score. No attempt was made to model insertions/deletions in the PSI-BLAST pseudo models, and part of the improvement achieved by the predictors can be traced to the modeling of insertions and deletions in parts of the alignment not covered by the template.

Improvement over Best Single Structural Template

At the core of template-based modeling lies the concept that a model for a protein can be constructed based on an alignment between a target protein sequence and a suitable template structure. In CASP7, we have introduced the concept to use a pseudo-model based on the single best available structural template as reference when measuring the quality of a prediction.[21] The actual predictions for a target are often worse than the reference model, e.g., in cases where only suboptimal templates were detected, or the target-template sequence alignment was imperfect. In other aspects, the predictions can significantly improve over the reference model, e.g., by successful *de-novo* modeling of unaligned regions, increasing coverage by multitemplate modeling, re-modeling structurally divergent segments, or successful model refinement.

As part of our assessment of the CASP9 TBM predictions, we wanted to determine to what extent the predic-

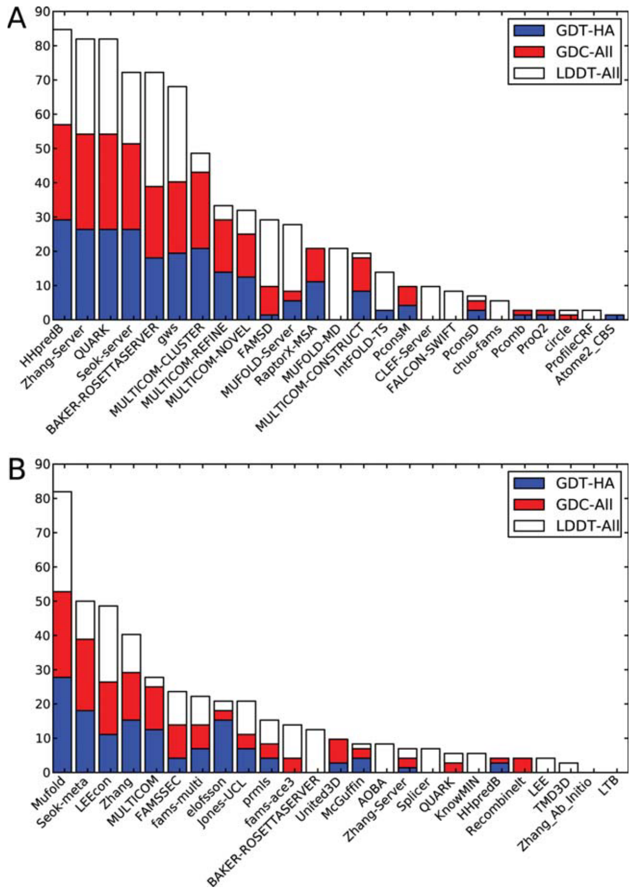


Figure 2

Ranking of the server (panel A) and human/server (panel B) categories based on a head-to-head comparison of the top 25 groups on common targets. In the "server" category the top groups have all similar performance, with group "449 HHPredB" standing out by a small margin. Conversely, group "386 Mufold" dominates the ranking in the "human/server" categories with a number of significant wins in the head-to-head comparison which is much larger than the one of its competitors.

tors were able to select suitable templates, and if they could generate models that approximated the target structure with a higher accuracy than simply copying the coordinates from the best available single template. For this analysis, the single best structural template (i.e., the template with the highest structural similarity) available in the PDB at the end of each target's prediction period was used as a reference. A raw backbone model based on the template was created for each target (See "Materials and Method" section for a detailed description of the procedure) and its GDT-HA score was compared with the scores of the predicted models. Figure 4 shows the result of this analysis (A for the "server" category and B for the "human/server" category). Each vertical bar in the plot represents a prediction group, with the fraction of targets predicted more accurately than the reference model proportional to the length of the bar in the positive half of the y-axis. The length of the bar on the negative side represents the fraction of targets for which the best prediction has lower accuracy than the template. No single group could generate models with at least the same accuracy of the best available template for the majority of the targets. The best group in the "server" category ("449 HHPredB") predicted 31.4% of the targets with higher accuracy than the single template reference model, while the best in the "human/server" category ("386 Mufold") achieved 34.5%. When the analysis shifts its focus from a single group to the whole range of predictions by all groups, the best prediction for each target, irrespective of its origin, had a higher GDT-HA score than the best available template in more than 60% of the cases for both the "server" category (77 of the 121 targets) and the "human/server" category (34 of the 55 targets). Accuracy improvements over the best available structural template observed in CASP9 are typically modest and mainly due to modeling of unaligned loop regions, insertions, and deletions, which are not represented in the reference models. The margin of improvement was in the vast

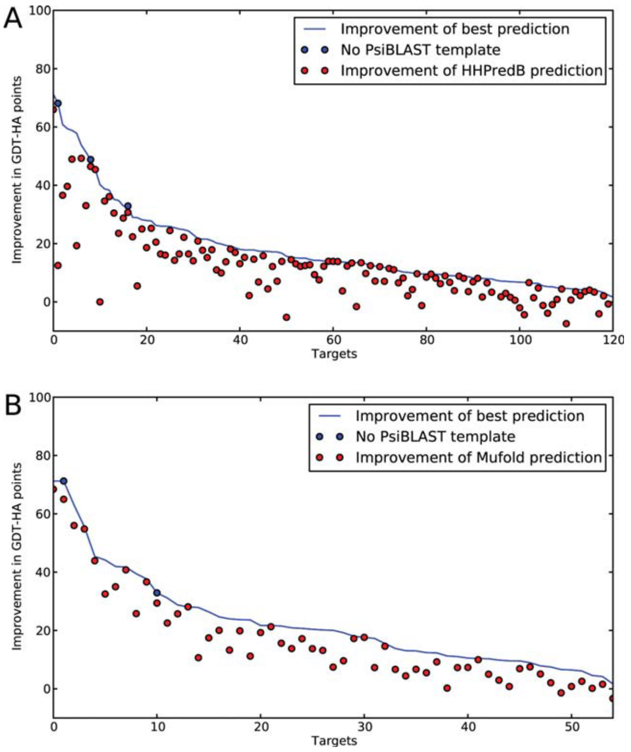


Figure 3

Overall CASP9 performance with respect to the best template detectable using PSI-BLAST, for the "server" category (A) and for "human/server" one (B). The continuous blue line shows the improvement in GDT-HA points that the best prediction for each target achieved. Targets for which PSI-BLAST returned no template are marked with blue circles. For comparison, the top ranked groups "449 HHPredB" for the "server" category and "386 Mufold" for the "human/server" category are indicated for each target as red dots. Targets are sorted by order of improvement.

majority of cases below ten GDT-HA units, with only a few exceptions: 19.0 GDT-HA units for target T0559, 22.1 for target T0580 (both by "321 BAKER-ROSETTASERVER"), and even 28.8 for target T0530 (by "228 YASARA") in the "server" category. In the "human/server" category, T0580 is the target with the largest improvement (25.72 GDT-HA units by "172 BAKER"), followed by T0619-D1 (19.30 points by "484 FEIG").

These results show that in the majority of cases, even the best groups could detect the best structural template for only a limited number of targets, and when they succeeded they could only in very few cases significantly improve their model beyond the information provided by the template. However, for almost two-thirds of the targets there was always at least one group that could identify the best template. This suggests that template-detection is still a field with room for improvement, especially when dealing with remote templates with very low sequence similarity to the target.

Quality of Template Selection and Target-Template Alignment

Since template detection and alignment between the target sequence and the template(s) are obviously limiting factors in template-based modeling, it would be interesting to independently assess these steps. However, in CASP only the final models enter the evaluation, with no information about the procedure how they were generated. It is therefore impossible to assess the individual steps of "template selection", "target-template alignment", and "model-building" independently of each other. The quality of the backbone is mainly influenced by choice of template and alignment, and of the three scores used during this assessment, GDT-HA is the one that does not depend on the quality of side-chain modeling and on the accuracy in modeling of local interactions. In an attempt to emphasize template selection and alignment accuracy, we selected a subset of the 50 hardest TBM targets as test set and used GDT-HA as a C_α only measure for the assessment. The best performance was achieved by the Raptor series ("77 RaptorX-MSA", "286 Raptor X", and "276 RaptorX-Boost"),[28] which stood out in the comparison on these challenging targets. Unfortunately, these methods did not stand out in the overall ranking due to limited accuracy of the all atom coordinates.

Prediction Example

Target T0580 (PDB: 3NBM) is used to illustrate a typical CASP9 prediction example. T0580 is the lactose-specific IIB component domain of the phosphoenolpyruvate carbohydrate phosphotransferase system (PTS) from *Streptococcus pneumonia*. This system transports and phosphorylates carbohydrates across the membrane at the expense of phosphoenolpyruvate (PEP) in many bacteria.[37] The best template identified by PSI-BLAST is another IIB component of the same system (*N,N'*-diacetylchitobiose) in *Escherichia coli* (PDB: 2WY2, chain D). The template covers 88% of the target sequence, with a sequence identity of 12.5%. A pseudo-model based on this template, as well as alternative at the time available structural templates, would have GDT-HA scores around 55. Nevertheless, this target was predicted well by a large number of groups, with the models from the best eight groups all having GDT-HA scores higher than 70. The absolute best predictions by human and server groups were by "172 BAKER" and "321 BAKER-ROSETTA-SERVER", respectively, with GDT-HA scores of 75.3 and 71.6. While among the human groups many almost matched the quality of the best prediction, lagging only 3–4 points behind, the prediction from the second best server groups had a GDT-HA score almost six points lower. When focusing the analysis of the performance on

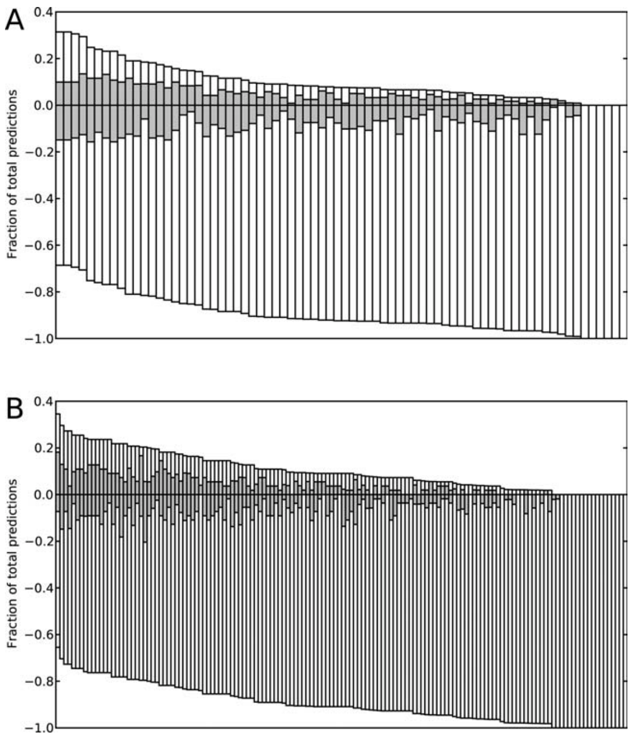


Figure 4

Performance of each group relative to the best single structural template available for each target. Panel A shows the data for the "server" category, panel B for "human/server", respectively. For each group, the fraction of total targets where the group improved over the best available template is plotted on the positive y-axis, while the fraction where the model was less accurate that the reference template is shown in the negative y-axis. The fractions are computed against the number of predicted targets (groups predicting less than half of the total number of targets have been excluded from the plot). Differences of less than two points in GDT-HA are plotted in a darker color shade. Even the best performing groups can predict models with higher accuracy than the best structural template only in roughly one-third of the cases.

the local scale, all the top groups modeled local atomic interaction with a lDDT-all score higher than 83. Once again, "172 BAKER" stands out by a small margin with an lDDT-all score of 89.8.[38] Figure 5 shows the model from the "172 BAKER" group (in red) and the template 2P4U (in blue) aligned to the target structure (in gray). The improvement in GDT-HA was mainly achieved by correctly modeling a long deletion in the template sequence and by correctly modeling the orientation of two helices that are skewed in the template structure. This results in a marked increase in GDT-HA score due to the relative small size of the target. T0580 represents a typical situation in CASP9, where structural templates were available for most targets and improvement over the template were mainly achieved by correct modeling of insertions and deletions.

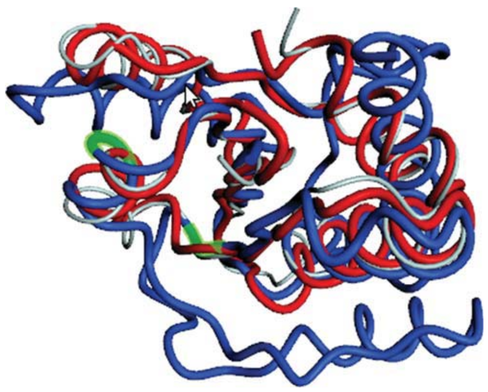


Figure 5

Example of successful modeling of target T0580. The best available structural template (PDB: 2P4U, chain B) is shown in blue. The best prediction ("172 BAKER") is shown in red. Both are aligned to the target structure in gray. The model has a GDT-HA score of 75.2, with an improvement of 25 points over the best structural template. This large improvement is achieved mainly by accurately modeling a long deletion in the template sequence, and by accurately predicting the orientation of two helixes which are skewed in the template structure.

Comparison between CASP8 and CASP9

Comparisons between different editions of CASP are difficult to perform because many variables change between different installments. Target difficulty is not easy to estimate. It is usually approximated by combining various parameters[22,39] and varies significantly between CASPs. Also the number of targets changed broadly across the last installments of CASP, from 108 in CASP7[21] to 154 in CASP8[19] to 121 in CASP9. Comparison of CASP8 and 9 with previous editions is further complicated by a change in target classification in "server" and "human/server" categories (with prediction periods of different length) after CASP7, and a resulting smaller number of targets with higher difficulty in the "human/server" category. Here, we present a limited comparison between CASP9 and CASP8 based on a relative criterion, namely the fraction of targets for which the predicted models showed a higher accuracy than the best available structural template. For a more detailed discussion about comparison between CASP9 accuracy with previous CASP editions, we refer to a dedicated manuscript elsewhere in this special issue.[40]

Pseudo-models using the best structural template for CASP8 targets available at the end of the prediction window (provided by the Prediction Center) were constructed using the same procedure described before for CASP9, and their GDT-HA score compared with the predictions. The average fraction of targets for which an improvement over the best structural template was observed (over the best ten groups) was then compared between CASP8 and CASP9 for both target categories (Fig. 6, A for the "server category" and B for the "human/server" category). Between CASP8 and CASP9 the fraction of targets showing improvement increased in the "server" category by a large margin across all difference thresholds. For the "human/server" category the improvement was less remarkable, probably due to a higher difficulty of the targets. In order to carry out a fair comparison, one should also take into account the relative difficulty of the targets in the two CASP experiments. For example, in CASP7 the best prediction group (Zhang) was reported to improve over the best structural template in about one-third of the cases, and the average fraction of improved targets for the best ten groups was similar to the value observed in CASP9.[21] It is unclear, which factors are responsible for the drop in performance witnessed in CASP8 according to this criterion, and obviously the various rounds of CASP pose different challenges to the predictors.

Estimation of Model Reliability ("Model B-Factor")

How useful would BLAST be without E-Values? Realistic estimates of the expected accuracy of a model are essential for any predictive method in order to be useful in a practical context. Since the usability of protein models for different applications in life sciences crucially depends on their accuracy, a realistic estimate of the expected model quality and coordinate errors is a prerequisite for practical application.[1] The assessment of several rounds of CASPs has demonstrated that the differences in accuracy between "easy" and "hard" targets are much larger than the differences between many of the participating methods for a specific target. From a user perspective, a reliable estimate of the expected overall accuracy of a model, and even more of individual residues and atom positions, is at least as important as the overall average modeling accuracy of a specific method.

In the CASP experiment, predictors are requested to provide confidence estimates for their predicted coordinates in the "B-Factor" column (expected error in Å). We have assessed the ability of the various prediction groups to estimate the accuracy of their own models using log-linear correlation between predicted and real deviations, as well as a ROC analysis of correctly identified errors. Again, as in previous CASP experiments, the accuracy of error estimates submitted by the majority of groups, including several of the top ranked ones, was overall very poor, and four prediction groups outperform all remaining groups in this respect by a large margin. According the area under curve (AUC) measure, the best classification of modeling errors was achieved by ”94 McGuffin”.[41] Three other predictors reach a similar level of accuracy "1 ProQ", "102 BILAB-ENABLE", and "275 IntFOLD-TS". Figure 7 shows the correlation between the local error estimates for the best group in this respect ("94 McGuffin") in comparison with the best server according to model accuracy in the TBM category ("449 HHpredB"). Successful predictors combine both accurate prediction of the coordinates with realistic estimates of model confidence and identification of modeling errors, thus providing additional value to the user by supplying hand in hand accurate coordinates and precise confidence values.

Evaluation of Oligomer Modeling

Many proteins form stable higher order quaternary structures in the form of complexes or oligomeric assemblies. In fact, proteins which form exclusively monomeric structures are a minority, while the majority of proteins in a cell is involved in complexes and assemblies in some form.[42] The prediction targets in CASP do not form an exception, and a large fraction of them is forming stable oligomeric assemblies in their native state: 53 prediction targets in the assessment were homo-oligomeric complexes, while only 43 were monomers.[‡] Frequently, it is only in the context of this assembly that we can understand their function, e.g., when active sites are located in the interface between subunits, the protomeric structure is often not sufficient to study functional aspects in an accurate and complete fashion. It is also often observed that the isolated subunits do not represent self-sufficient globular structures without the interactions with their neighboring subunits. In these cases, predicting (and also assessing) targets as monomeric structures does not make much sense.

The task of predicting protein structures in CASP includes prediction of the quaternary structure. According to the CASP format definition, quaternary structure predictions should be submitted in the same frame of reference, with the first chain labeled as A and subsequent chains following the Latin alphabet, e.g., a tetramer's chains should be labeled as A, B, C, D. Here, we evaluated how well predictors identified the correct oligomeric state of the target, and how accurately the predictions resembled the structure of the complexes. To be able to estimate the difficulty of the different targets, two naïve predictors were included in the assessment: group "998 NaïveSeqId" and group "999 NaïveCoverage". These naïve predictors assume that the oligomeric state of the target is the same as the one of the closest template which can be identified by standard sequence search methods. In case of group "998 NaïveSeqId", the template with the highest sequence identity to the target was selected, in case of group "999 NaïveCoverage", the one covering the largest fraction of the target sequence.

The oligomeric state of CASP9 targets ranges from "Monomeric" to "Tetrameric". The majority of targets is multimeric, with "Dimer" as the most abundant state is (41 Targets), followed by "Tetramer" (9) and "Trimer" (3). The "human/server" sub set consists of 20 monomers, 16 dimers, 3 trimers, and 2 tetramers.

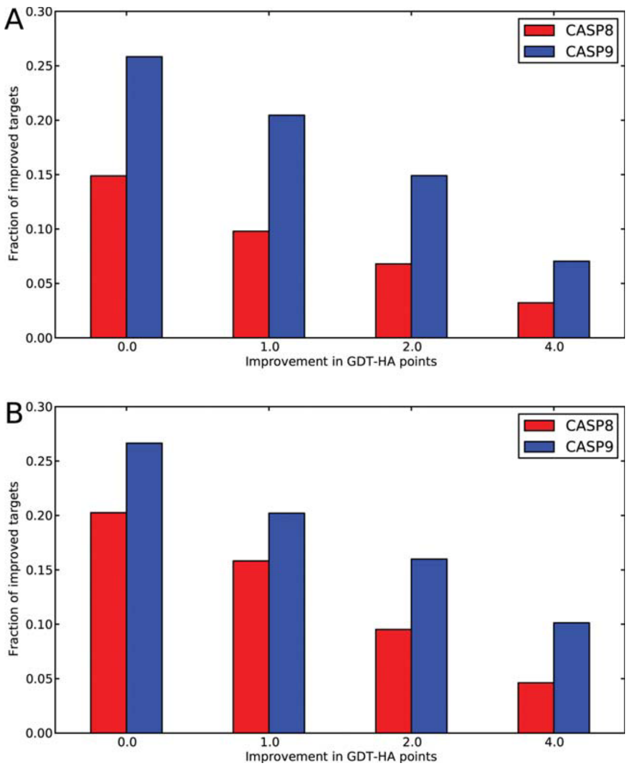
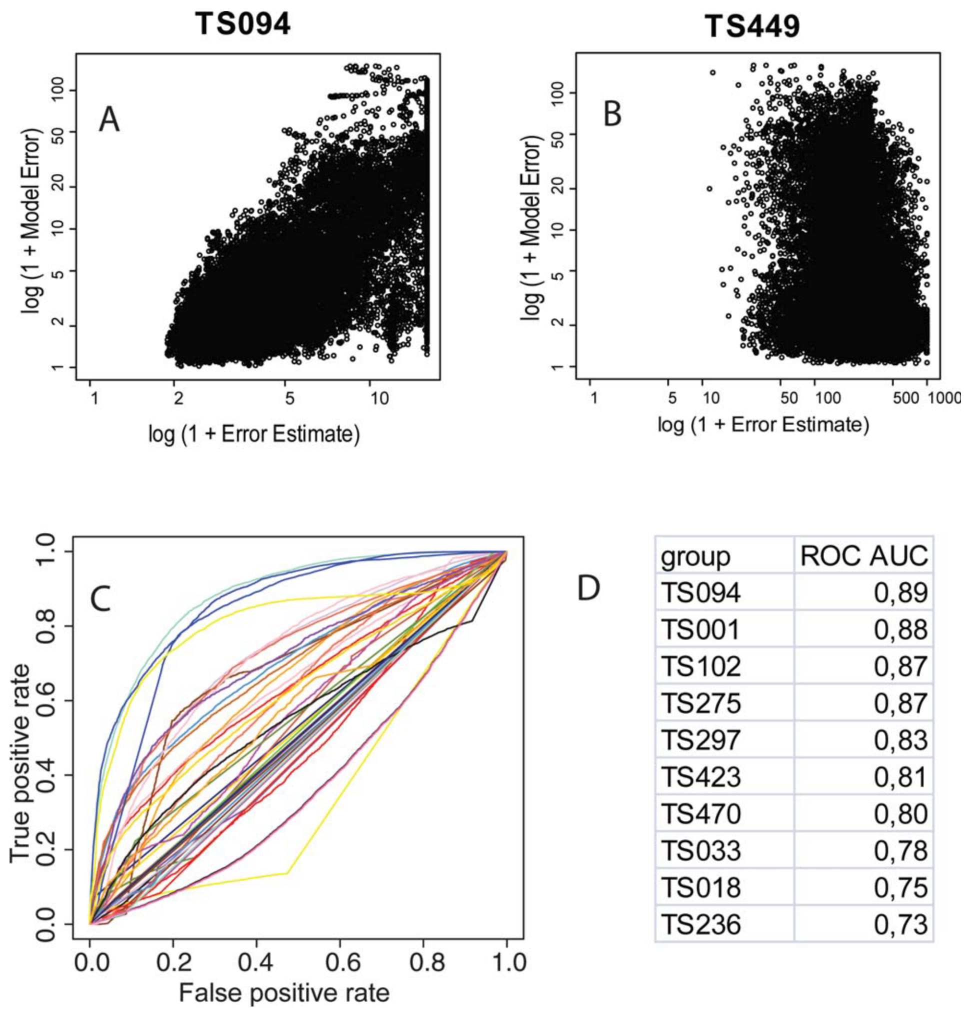


Figure 6

Fractions of targets for which an improvement over the best available structural template could be observed in CASP9 (blue) and CASP8 (red). Panel A shows the data for the "server" category, while panel B refers to the "human/server" one. The left-most group of bars shows the fraction of targets for which any improvement was achieved. The other groups refer to improvements of at least 1.0, 2.0, and 4.0 GDT-HA units, respectively. A clear advancement is visible between CASP8 and CASP9 in both categories.

[3] ‡Some targets, for which the oligomeric state assignment by the experimentalist was ambiguous, were excluded from this part of the assessment.



group	ROC AUC
TS094	0,89
TS001	0,88
TS102	0,87
TS275	0,87
TS297	0,83
TS423	0,81
TS470	0,80
TS033	0,78
TS018	0,75
TS236	0,73

Figure 7

Assessment of model confidence estimates (“Model B-Factors”). Panels A and B show the correlation between the local error estimates for all residues for the best group in this respect (“94 McGuffin”) in comparison with the best server according to model accuracy in the TBM category (“449 HHpredB”), respectively. According to the ROC analysis, four groups provide significantly better local error estimates of their own predictions than all other groups (panels C and D).

We included all groups in the evaluation which submitted at least one model as oligomer. The majority of targets was predicted as monomeric (83% “human/server”, 78% “server” category), followed by dimeric predictions (13% “human/server”, 18% “server”). This indicates that many of the groups only submitted a fraction of their predictions as oligomers. Only for a small number of groups does the oligomeric state distribution of the predictions resemble the one of the experimental target structures [Figure 8 (A,B)].

Successful modeling of oligomeric complexes relies on the identification of the correct oligomeric state of the unknown target protein, and then on constructing a realistic model for the quaternary structure. The fraction of correctly identified oligomer states over the number of submitted models ($\mathrm{Acc}_{\mathrm{Rel}}$ column in Table III) ranges between 79% (“282 Taylor”) and 36% (“20 dokhlab”) for “human” and 66% (“452 Seok-server”) to 45% (“102 Bilab-ENABLE”) for “server groups”, respectively. However, these numbers are dominated by the assignment of monomeric structures. In the context of CASP, it is not possible to determine if a monomeric submission was an explicit choice for monomer or just a “non-oligomeric” prediction. (If a predictor would submit all models as monomeric, an accuracy of 47% and 53% would be achieved for “human/server” and all targets, respectively.) Also, submitting oligomeric predictions only for a small set of targets (“cherry picking”) increases the chance to achieve good accuracy in this measure.

We therefore calculated the fraction of correctly predicted states for oligomeric targets only, normalizing by the maximum number of oligomeric structures, either in the target or in the prediction set (See Materials and Methods

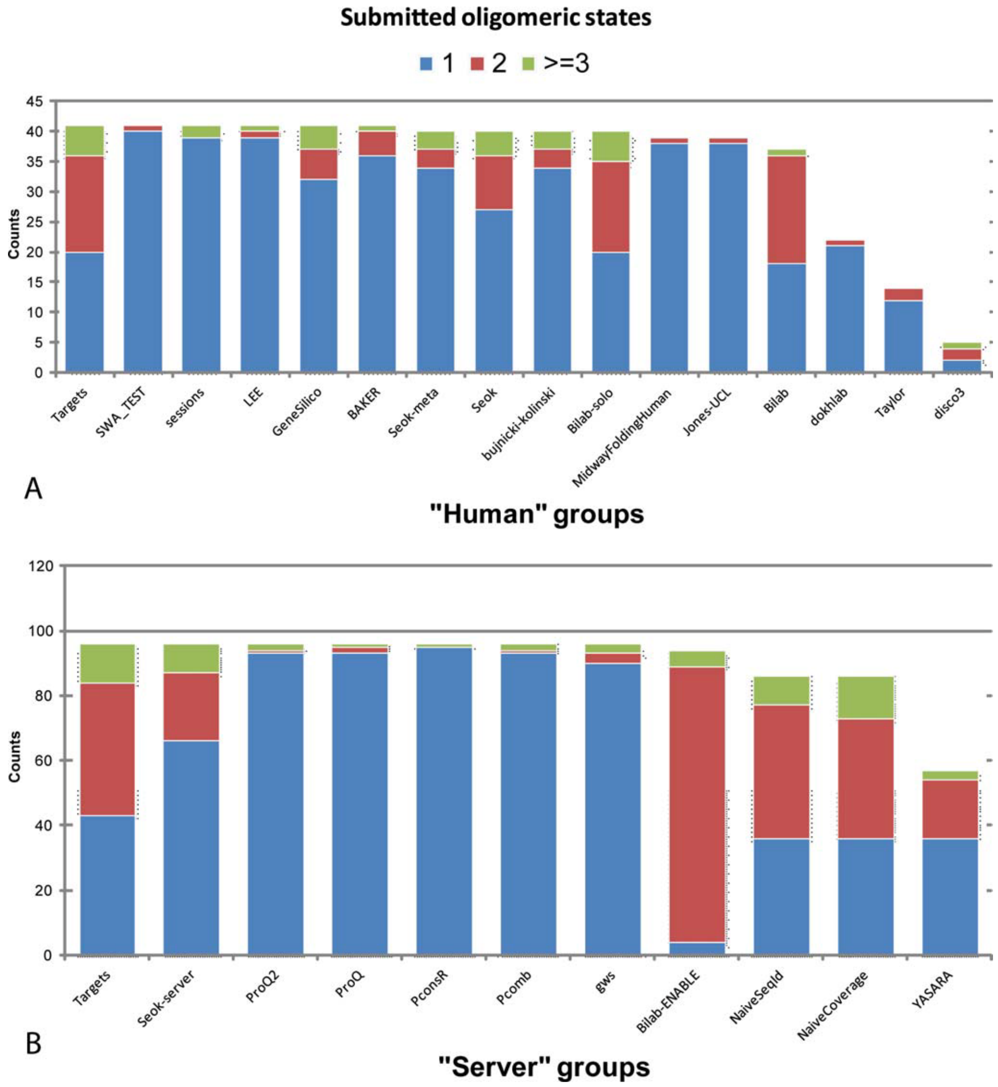


Figure 8

Distribution of oligomeric states among the submitted models. Panel A shows the fraction of different oligomeric states in the set of predictions made by "human" groups for the targets in the "human/server" category. Panel B displays the same data for the set of the submissions made by "server" groups for all evaluated targets. Both panels show on the left the actual distribution of the oligomeric states in the experimental target structures for comparison (see text).

for details). Two human groups ("458 Bilab-solo", "242 Seok") and the two naïve predictors predicted more than 50% of their oligomeric targets correctly. Remarkably, group "458 Bilab-solo" was able to characterize more than three quarters (76%) of the oligomeric states in the "human/server" targets correctly. In the "server" category, the naïve predictor based on sequence identity, classified most accurately 55% of the targets (See Table III).

In our visual inspection of the predictions, we observed a substantial fraction of physically impossible models with parts of the structures overlapping, severe steric clashes, or isolated chains in space—lacking a protein–protein interface. We calculated the fraction of models which contained more than 10 backbone-backbone clashes or were lacking inter-chain contacts (all C_β–C_β distances > 12 Å). In general, server groups tended to build more unrealistic models than human groups. Among the groups having at least five oligomeric predictions, the server group "102 Bilab-ENABLE" had with 36% the highest fraction of unrealistic models (See Table III).

In order to characterize the accuracy of the predicted oligomeric structures, we calculated a "Contact Agreement Score" S_{agree} which reflects the fraction of correctly modeled interface contacts in the complex and thereby accounts for the correct number of interfaces as well as their correct orientation (see Materials and Methods). S_{agree}–score ranges from 1 for a perfectly predicted complex to 0 for a completely incorrect one. To estimate the

Table III
Summary Table of Oligomeric Predictions

Category	Group	Group#	$\sum$oligomer predictions	Total	$\mathrm{Acc}_{\mathrm{Rel}}$	$\mathrm{Acc}_{\mathrm{Oli}}$	% unrealistic	$\sum S_{Agree}$
Human	Seok-meta	16	6	40	57.5%	23.8%	0.0%	1.2
Human	dokhlab	20	1	22	36.4%	0.0%	0.0%	0.0
Human	Jones-UCL	104	1	39	53.8%	4.8%	0.0%	0.0
Human	LEE	114	2	41	51.2%	4.8%	20.0%	0.6
Human	GeneSilico	147	9	41	61.0%	28.6%	11.1%	2.3
Human	BAKER	172	5	41	61.0%	23.8%	0.0%	1.8
Human	Seok	242	13	40	75.0%	57.1%	7.7%	3.1
Human	sessions	278	2	41	51.2%	4.8%	0.0%	0.4
Human	Taylor	282	2	14	78.6%	9.5%	0.0%	1.3
Human	SWA_TEST	297	1	41	51.2%	4.8%	0.0%	0.8
Human	bujnicki-kolinski	299	6	40	62.5%	23.8%	0.0%	2.2
Human	Bilab	423	19	37	51.4%	42.9%	15.8%	1.3
Human	disco3	439	3	5	60.0%	9.5%	0.0%	1.0
Human	Bilab-solo	458	20	40	77.5%	76.2%	0.0%	4.9
Human	MidwayFoldingHuman	477	1	39	48.7%	4.8%	0.0%	0.6
Server	ProQ2	1	3	96	46.9%	3.8%	33.3%	0.9
Server	Bilab-ENABLE	102	90	94	44.7%	45.6%	35.6%	7.6
Server	PconsR	173	1	96	44.8%	0.0%	100.0%	0.1
Server	YASARA	228	21	57	63.2%	28.3%	9.5%	5.1
Server	gws	236	6	96	50.0%	9.4%	33.3%	1.9
Server	Pcomb	273	3	96	46.9%	3.8%	33.3%	1.0
Server	ProQ	296	3	96	46.9%	3.8%	0.0%	0.8
Server	Seok-server	452	30	96	65.6%	43.4%	6.7%	6.4
Server	NaiveSeqId	998	50	86	60.5%	54.7%	20.0%	8.1
Server	NaiveCoverage	999	50	86	57.0%	50.9%	8.0%	9.8

overall performance for each group, we summed up the S_{agree}–score for each target. This procedure rewarded successful modeling of a complex but did not penalize unsuccessful attempts, which also accounts for the fact that is was often not clear if a "single chain" submission should be considered as explicit choice for a monomer or no attempt was made to predict the oligomeric state. When considering the "human/server" subset of targets, group "458 Bilab-solo" submitted overall the most accurate predictions (see Figure 9, $\sum S_{agree} = 4.9$). By evaluating the server groups on all targets, the naïve predictor which uses the template with best coverage was the most accurate ("999 NaïveCoverage"), followed by the second naïve predictor ("998 NaïveSeqId"). To investigate if human predictors were able to submit more accurate models than servers, we compared server and human groups on the subset of "human/server" targets. Among the best five groups there were two human predictors, two server predictors and the naïve predictor "999 NaïveCoverage". The human group "458 Bilab-Solo" outperformed all other groups, but a general trend that human groups predict more accurately could not be observed. For most targets (except T0584, T0517, T0598), "458 Bilab-solo" predictions achieved an accuracy similar to the top predicting groups and thus performed on a constant level (Supporting Information Fig. SIII).

Despite the fact that the quaternary structure is often essential for the understanding of the biological function of a protein and more than half of all CASP9 targets are oligomers, only a small fraction of the participating groups in CASP9 submitted oligomeric models. The overall performance of the participating server groups compared to the two naïve predictors was rather poor. As shown in Figure 8(B), most of the sever groups submitted mainly monomeric submissions, except group "102 Bilab-ENABLE", who submitted mainly oligomeric models. Unfortunately, a significant part of these models contained clashes or have non-interacting subunits. In general, the accuracy of the server oligomer predictions appeared rather low. The high rate of unrealistic models reveals that the complexity of oligomeric modeling is currently not handled properly by automated approaches. Obviously, there is a great opportunity for significant improvement in modeling of quaternary structures of proteins in future CASPs.

Example of an Oligomer Target

Ignoring the quaternary structure of a protein can lead to models which cannot explain important physiological properties, or even to structures with a disrupted functional site. One example is T0576 (PDB: 3NA2), a functionally uncharacterized protein from *Leptospirillum rubarum*. This target shows in its oligomeric form an extensive interaction network at the interface between the monomers. If only the monomeric structure is considered, one beta sheet remains exposed, extending into the solvent in a situation that is obviously not energetically favorable. In the dimeric form, sheet pairing causes the exposed hydrophobic residues to become buried (Figure 10). A homologue structure of this target (a putative heme-binding protein from *Anabena variabilis*—PDB: 3FM2) shows a very similar configuration of the binding site and includes

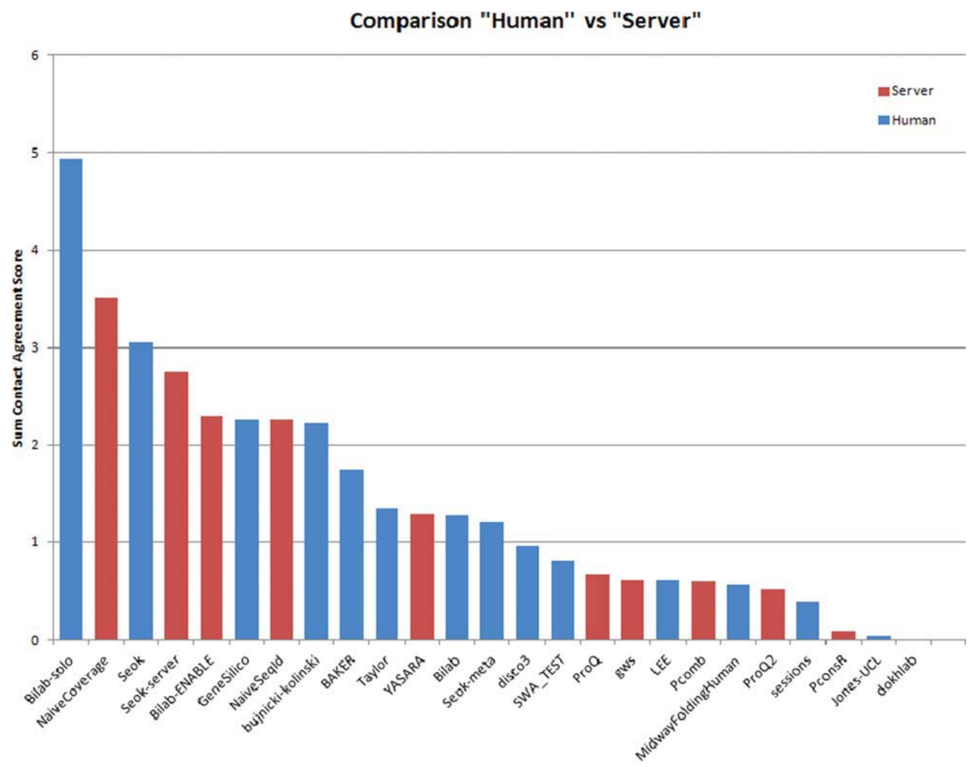


Figure 9

Interface agreement scores S_{agree} summed over all targets in the “human/server” category. “458 Bilab-solo” clearly outperforms all other groups. In general, predictions by “human” predictors (blue) are not more accurate than “server” groups (red) and only “458” outperforms a naïve control predictor in this analysis.

two zinc ions located in the clefts formed by the two chains. This finding supports the hypothesis that the dimeric form of target T0576 is a requirement for its biological function.

The group that submitted the best prediction for this target, “458 BILAB-solo” ($S_{agree} = 0.79$), was successful in modeling the interface region, including the sheet pairing. Almost all inter-chain contacts present in the interface region of the target have been correctly modeled in the prediction. For biological applications, the usefulness of a correct oligomeric model like the one discussed here, which shows the protein in what is likely to be its functional state, clearly exceeds the one of any monomeric model.

Quality of Binding Site Modeling

In the context of biological relevance, correct modeling of binding sites is an integral part. Within CASP9 30 targets were found to contain ligands related to their biological function. While the quality of the binding site is implicitly included in the overall TBM assessment, we evaluated this aspect in more detail by comparing the overall backbone accuracy of the domain containing the binding site to the local accuracy of only the binding site residues. The binding sites residues were identified by the same strategy outlined in the report on the assessment of the ligand binding residue predictions.[43] Not surprisingly, overall low quality of the whole structure is accompanied by a low quality of the binding site (Supporting Information Fig. SII). However, with respect to the biological relevance of the models we focused the analysis on high quality models with at least 60 GDT-HA for the ligand binding domain. For individual targets, the quality of the binding site model varies considerably and no correlation between backbone quality and active site accuracy can be observed [Figure 11(A)]. Some groups submitting a good overall quality backbone model, often only built an average quality binding site. Vice versa, some groups were modeling some binding sites rather well, but still delivered only average overall models. In particular the former finding is illustrated in Figure 11 (B) for target T0632 (PDB: 3NWZ). The protein is a putative thioesterase in *Bacillus halodurans*, which is a homo-tetramer, binding co-enzyme A where the ligand is interacting with three of the four chains of the protein. In this case, group “361 LeeCon” provided a good model for the binding site (lDDT 82.3), whereas group “63 JIANG_ASSEMBLY” submitted a model containing a clash at residue TYR117 within the otherwise well modeled binding site (lDDT 72.1). Both groups obtained reasonable scores for the ligand binding domain (GDT-HA 76.1 for group 361 and 75.4 for group 63). In summary

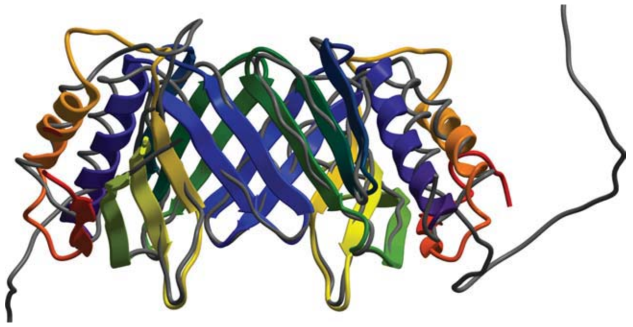


Figure 10

T0576—example for a dimeric prediction target in CASP9 (group "458 Bilab-solo". An isolated monomeric subunit would not form a self-contained structure and leave several of the β-sheet interactions unsatisfied.

many targets show a considerable range in quality concerning the ligand binding site, which indicates room for improvement, especially since this aspect is crucial for the biological relevance of the models.

MATERIALS AND METHODS

Target Structure Processing and Assessment Unit Definition

The procedure for target structure classification and assignment of assessment units (AU) is described in detail in a separate paper in this special issue.[12] Here we briefly summarize aspects relevant for the TBM assessment. For each experimental target structure, a single chain was selected as reference. When several chains with identical sequence were available, the chain with the highest coverage was selected. All available chains in the experimental structures were superposed (using the program TM-Score[44]) to identify segments with extensive structural variability. Stretches with C_α of equivalent residues deviating by more than 3.5 Å in different chains were classified as flexible and excluded from the AU. Finally, each structure was analyzed in the context of the crystallographic unit cell observed during structure determination (when available) to exclude loop regions with obvious crystal contact artifacts. Target structures were subdivided into separate assessment units (AUs) if (a) there was clear indication for significant domain flexibility and inter-domain motions in the target structure or the submitted predictions, and (b) if the two AUs would be assigned to different categories (FM or TBM, respectively). A discussion of the procedure can be found in Shi *et al.*[26] The target preparation process was carried out using a consensus procedure, with assessors reviewing and confirming each other's findings until an agreement was reached.

Numerical Scores

GDT and GDC scores were computed with the latest version of LGA.[13] Manipulations of target structures and predictions were implemented using the OpenStructure framework.[45] For the assessment of close inter-atomic contacts ("van der Waals clashes"), a list of short distance contacts observed in 1769 high resolution protein structures was derived. All entries shared no more than 20%

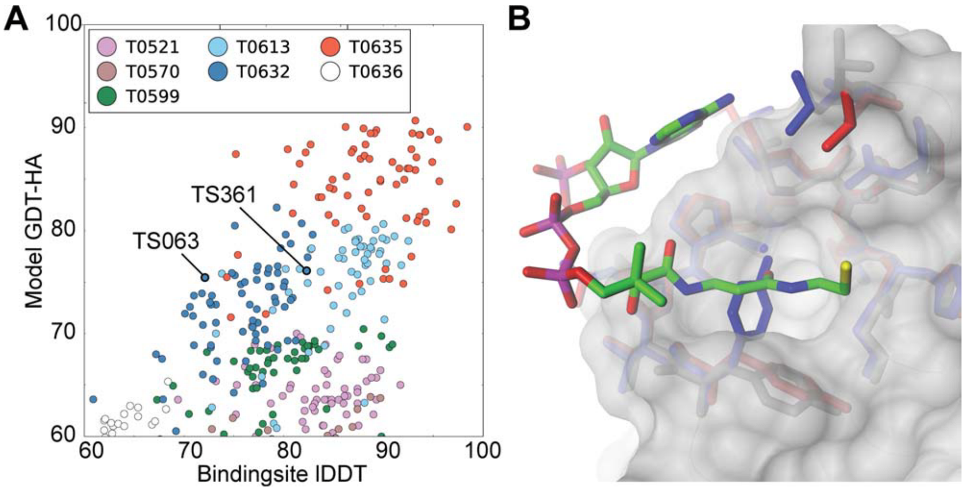


Figure 11

Does backbone accuracy correlate with the accuracy of binding sites? Panel A shows comparisons of the binding site lDDT-all scores versus the GDT-HA scores for targets with an average GDT-HA of at least 50. The quality of the binding sites varies considerably—even for models with comparable backbone accuracy. Panel B shows an example of two models for target T0632. The binding site is only partially shown for visual clarity, binding site residues are displayed in sticks and the target structure and surface are shown in gray. A good overall model (group "63 JIANG_ASSEMBLY" in blue) contains a clash with the ligand in the binding site at TYR117, whereas group "361 LeeCon" (in red) submitted an equally good backbone model, with a well modeled binding site.

sequence identity and their structures had been determined to a resolution of at least 1.6 Å and with an *R*-free value of 0.25 or less. The non-redundant set was created using the PISCES server.[46] A distance distribution for each different pair of non-bonded atoms was plotted to define the minimal distance thresholds. Two non-bonded atoms in a prediction were considered "clashing" if their positions were closer than their element specific reference distance of C:C 1.9, N:N 2.1, O:O 1.8, S:S 1.4, C:N 1.9, C:S 2.4, C:O 2.1, N:S 2.1, N:O 1.7, O:S 2.1 (in Å).

Local Distance Difference Test on All Atoms (lDDT-all)

The local Distance Difference Test score was computed using the following procedure: a list of pair-wise non-bonded distances was generated from the target protein structure. For each atom *i*, all atoms *j* not part of the same residue as *i* and lying within 5 Å from *i* were considered as interactions partners of *i*. The cumulative list of *i*-*j* interactions stemming from all atoms in the experimental protein structure was taken as reference against which to score predictions. Specifically, interaction distances in the protein structure were compared with distances between corresponding atoms in the predictions. If the difference between the two distances was below a defined threshold, the interaction was considered to be preserved in the prediction. The final lDDT-all score was computed by averaging the fraction of correctly modeled interactions for the following four distance difference thresholds: 0.5, 1, 2, and 4 Å (the same thresholds as GDT-HA). It must be noted that residues with ambiguous nomenclature for chemically equivalent atoms, for which the choice of atom nomenclature could influence the final score, were dealt with by computing a score using all possible nomenclatures, and by choosing the one giving the highest value.

Residues considered physically or chemically implausible (e.g., side-chains involved in at least one serious van der Waals clash) were considered as modeled incorrectly and all interactions involving atoms of this specific side-chain were considered incorrect. In cases, where backbone atoms were involved in serious van der Waals clashes, all interactions involving atoms of this specific residue were considered incorrect. Obviously, distances which would involve non-predicted or incomplete residues were considered incorrect too. The Distance Difference Test calculation was implemented using the OpenStructure framework.[45]

Best PSI-BLAST Template

Using the sequence of each target PSI-BLAST was used to search for suitable templates. Sequence profiles were generated using five iterations of PSI-BLAST against a snapshot of the NR database at the time of the CASP9 experiment clustered at 90% sequence identity.[14] These profiles were then used to perform a search against the PDB database.[47] PSI-BLAST hits were filtered according to the release date to remove the template structures which were not available during the prediction period of each target. The hit with the highest coverage of each separate AU was then chosen as the template for that specific AU. Using the alignment provided by PSI-BLAST, models were generated by copying the backbone coordinates of aligned residues. The GDT-HA scores of these models were then compared with the scores of the predictions.

Best Single Structural Template Models

At the end of each target's prediction window, the Prediction Center released a list of the best structural template available for each Assessment Unit, ranked by LGA_S score.[13] The list provided by the Prediction Center was extended by all structures in the PDB[47] database within 90% sequence identity, and an optimal structural alignment was generated for all possible templates by aligning them to the target structure using a 4 Å LGA sequence-independent superposition. Based on these alignments, backbone pseudo models were built by copying the backbone coordinates for aligned residues. Insertion and deletions were ignored, with no attempt to model them. The result of this process was a pool of models reflecting the best structural templates available at the end of the target's prediction period. The models were sorted according to GDT-HA; the model with the highest GDT-HA score was then selected as the reference model for respective target AU under investigation.

Assessment of Model Confidence Values ("B-Factor")

A residue in a prediction was considered modeled correctly if its C_α distance was not further than 3.5 Å apart from the experimental structure after global LGA superposition. Confidence values ("B-Factor") were re-ranked between 0 and 1 and the enrichment of correctly identified model errors was plotted under different thresholds by an ROC analysis. The area under curve ("AUC") was used as measure for the ability to assign realistic error estimates.

Assessment of Oligomeric Assemblies

Target preparation

For all CASP9 TBM targets, we determined the most probable biological active quaternary structure in the following way: for the definition of the oligomeric assembly state, we relied primarily on the assignment by the authors ("REMARK 350"). For targets solved by NMR, having no "REMARK 350" section, the oligomeric state was defined by their assembly of chains in the PDB entry. Targets without or with ambiguous assignments by authors were inspected manually taking into account PISA annotation[48] and the "REMARK 300" section. Targets with ambiguous assignment of the oligomeric state which could not be resolved satisfactorily by visual inspection were excluded from the evaluation. For two targets, the structure was not yet deposited in the PDB. Supporting Information Table SI provides the oligomeric state assignment for all targets used in this assessment.

Coordinate sets representing the biological units were downloaded from the PDB protein database or PISA respectively using the PDB code for the targets reported on the CASP9 target website. Residues in the experimental structure of the oligomeric assembly were mapped to the CASP target sequence chain-by-chain, and only amino acid residues corresponding to the CASP target sequence were included.

Oligomeric predictions

TS predictions were considered as oligomer predictions if a model consisted of multiple chains, and the oligomeric state of a prediction was interpreted as the number of chains found in the corresponding coordinate file submitted to the prediction center. Groups with at least one oligomeric model submission were included in the evaluation. Groups "55 MUFOLD-MD", "117 3-D JIGSAW_V4-5", and "333 DELCLAB" submitted models with inconsistent chain naming and were therefore excluded from this evaluation. Human groups were evaluated using the targets labeled as "human/server", server groups on all targets. Group "353 SAMUDRALA" (registered as "human") submitted in total only one oligomeric prediction (T0516) which is classified as "server" and was therefore not included in the assessment.

Numerical oligomeric state assessment

We calculated the fraction of correctly predicted oligomeric states (dimer, trimer, tetramer, etc.) by normalizing with the maximum number of oligomeric structures, either in the target or in the prediction set, in order to account for over-prediction of oligomeric states:

$$\mathrm{Acc}_{\mathrm{Rel}} = \frac{\text{number of correctly predicted multimeric targets}}{\max(\text{number of multimeric targets, number of multimeric predictions})}$$

In order to assess the quality of the structure of the predicted complex, we evaluated how accurately the interfaces of the oligomeric complexes were modeled. This analysis accounts for the correct number of interfaces as well as the correct orientation by calculating a "Contact Agreement Score" S_{agree} which reflects the fraction of correctly modeled interface contacts in the complex:

$$S_{\mathrm{agree}} = \frac{\sum_{i,j} f(x_{ij}, y_{ij})}{\sum_{i,j} g(x_{ij}, y_{ij})}$$

$$f(x_{ij}, y_{ij}) = \begin{cases} 1 - \frac{|x_{ij} - y_{ij}|}{\max(x_{ij}, y_{ij})}, & \max(x_{ij}, y_{ij}) > 0 \\ 0, & \max(x_{ij}, y_{ij}) = 0 \end{cases}$$

$$g(x_{ij}, y_{ij}) = \begin{cases} 1, & \max(x_{ij}, y_{ij}) > 0 \\ 0, & \max(x_{ij}, y_{ij}) = 0 \end{cases}$$

Two residues were considered to be in contact if they were located in different chains and their corresponding C_β atoms (C_α for GLY) were less than 12 Å apart. x_{ij} (y_{ij}) are the absolute number of contacts between residue i and j in the target (model) structure. Indices i and j represent the position of the residues in the sequence of the CASP target monomer (e.g., a tetrameric protein has four residues with index i).

S_{agree} ranges from 0 and 1, with $S_{\mathrm{agree}} = 1$ indicating that all contacts in the target complex are present in the model. $S_{\mathrm{agree}} = 0$ indicates that none of the contacts in the target complex are present in the model. Note that S_{agree} is undefined if either one of the two compared structures is monomeric and was set to 0 in this case.

Naïve oligomeric assembly predictors

To be able to estimate the difficulty of the different targets, two naïve predictors were included in the assessment: group "998 NaïveSeqId", and group "999 NaïveCoverage". HHSearch[33] was used to identify homologue template structures in the PDB, and all hits showing sequence identity with the target of less than 15% or coverage less than 15% were discarded. Group "998 NaïveSeqId" sorted possible template first by highest sequence identity in the target-template alignment and second by coverage of the target sequence by the alignment, giving precedence to the first criterion and selected the highest ranked template. Conversely, group "999 NaïveCoverage" selected the model, where its target-template alignment reached the highest coverage to the target sequence, and within the same coverage had the highest sequence identity, in this order of importance.

For the selected template, the first oligomeric assembly assigned by PISA[48] was used to build oligomeric pseudo-models. Models were built by copying the backbone atoms and the C_β atoms (except for Glycine) of the aligned regions in the sequence alignment.

CONCLUSIONS

The assessment of the predictions in the category of template based modeling in CASP9 has confirmed several trends which were already observed in previous years: fully automated prediction servers continue to play an increasingly important role. The category of "human" predictor groups was again dominated by meta-methods, which base their predictions on server models as input. We found no indication that human structural expert knowledge has played a significant part in the prediction process itself in the form of "individual expert prediction". Also the results of group 170 FOLDIT—the Foldit multiplayer online game[49]—could not demonstrate that a large number of participants in form of "crowd-sourced" human-directed computing were able to collaboratively address the challenges in protein structure prediction in the context of CASP9.

Template detection methods have clearly advanced far beyond PSI-BLAST and continuously improve in identifying the best available structural templates also for cases of remote homology. As in previous years, the main focus of the predictor groups was on achieving highest accuracy on isolated domain structures, and only little effort was made by most groups to model quaternary structures, or to provide realistic confidence estimates.

During the CASP9 evaluation, we strongly emphasized the assessment of all-atom models, including the quality of side chain conformation, and the accuracy of quaternary structures. In our view, this trend should be continued in future editions of CASP for two reasons: first, the template-based modeling field has matured to a level where it is reasonable to require correct modeling of local interactions. Second, for most biological applications, accurate backbone models of isolated domains are usually not sufficient, and all-atoms models of the correct functional state of a protein are required.

During the CASP9 meeting in Asilomar, it became apparent that the current form of the assessment has limitations in addressing the needs of the different communities: from a developer perspective, it would be beneficial to be able to quantitatively benchmark individual steps of the modeling process, e.g., template selection, target-template alignment, or loop building. However, this is not possible in the current setup as assessing the final model in most cases does not allow disentangling the influence of the different steps on the final outcome. On the other side, from an end user perspective the accuracy of the final models, including aspects such as correct quaternary structure and reliable error estimates, is of highest relevance. However, for many groups participating in CASP9 development time is limited and it is hardly possible to develop own methods for all different aspects of the modeling process.

Two alternative ways for solving this dilemma have been suggested during the meeting: the CASP process could be subdivided artificially into assessing individual steps separately, or different groups could collaborate more closely to form stronger prediction teams by combining their complementary expertise. The second option clearly appears more attractive from an end user perspective.

ACKNOWLEDGMENTS

We thank all experimental groups and participating predictors, without whom CASP would not be possible. We wish to thank our fellow assessor Nick Grishin and his team for the constructive collaboration, and the Prediction Center, especially Andriy Krystafovych, for handling the submission procedures and providing data for the analysis. Last but not least, we thank John Moult and Anna Tramontano for continuous encouragement and valuable discussions. We gratefully acknowledge support of our group by the SIB Swiss Institute of Bioinformatics.

REFERENCES

1. Schwede T, Sali A, Honig B, Levitt M, Berman HM, Jones D, Brenner SE, Burley SK, Das R, Dokholyan NV, Dunbrack RL, Jr., Fidelis K, Fiser A, Godzik A, Huang YJ, Humblet C, Jacobson MP, Joachimiak A, Krystek SR, Jr., Kortemme T, Kryshtafovych A, Montelione GT, Moult J, Murray D, Sanchez R, Sosnick TR, Standley DM, Stouch T, Vajda S, Vasquez M, Westbrook JD, Wilson IA. Outcome of a workshop on applications of protein models in biomedical research. Structure 2009;17:151–159.
2. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 2011;39:D225–229.
3. de Lima Morais DA, Fang H, Rackham OJ, Wilson D, Pethica R, Chothia C, Gough J. SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic Acids Res 2011;39:D427–434.
4. Lees J, Yeats C, Redfern O, Clegg A, Orengo C. Gene3D: merging structure and function for a Thousand genomes. Nucleic Acids Res 2010;38:D296–300.
5. DiMaio F, Terwilliger TC, Read RJ, Wlodawer A, Oberdorfer G, Wagner U, Valkov E, Alon A, Fass D, Axelrod HL, Das D, Vorobiev SM, Iwai H, Pokkuluri PR, Baker D. Improved molecular replacement by density- and energy-guided protein structure optimization. Nature 2011;473:540–543.
6. Schwarzenbacher R, Godzik A, Jaroszewski L. The JCSG MR pipeline: optimized alignments, multiple models and parallel searches. Acta Crystallogr D Biol Crystallogr 2008;64:133–140.
7. Kalyanaraman C, Imker HJ, Fedorov AA, Fedorov EV, Glasner ME, Babbitt PC, Almo SC, Gerlt JA, Jacobson MP. Discovery of a dipeptide epimerase enzymatic function guided by homology modeling and virtual screening. Structure 2008;16:1668–1677.
8. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch EM, Wilson IA, Baker D. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science 2011;332:816–821.
9. Goodsell DS, Olson AJ. Structural symmetry and protein function. Annu Rev Biophys Biomol Struct 2000;29:105–153.
10. Marianayagam NJ, Sunde M, Matthews JM. The power of two: protein dimerization in biology. Trends Biochem Sci 2004;29:618–625.

11. Kryshtafovych A, *et al.* Target Highlights in CASP9: Experimental Target Structures for the Critical Assessment of Techniques for Protein Structure Prediction. Proteins 2011;79(Suppl 10):6–20.
12. Kinch LN, Shi S, Cheng H, Cong Q, Pei J, Mariani V, Schwede T, Grishin NV. CASP9 Target Classification. Proteins 2011;79(Suppl 10):21–36.
13. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
14. 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.
15. Kryshtafovych A, Fidelis K, Tramontano A. Evaluation of model quality predictions in CASP9. Proteins 2011;79(Suppl 10):91–106.
16. Clarke ND, Ezkurdia I, Kopp J, Read RJ, Schwede T, Tress M. Domain definition and target classification for CASP7. Proteins 2007;69 Suppl 8:10–18.
17. Tress ML, Ezkurdia I, Richardson JS. Target domain definition and classification in CASP8. Proteins 2009;77 Suppl 9:10–17.
18. Battey JN, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T. Automated server predictions in CASP7. Proteins 2007;69 Suppl 8:68–82.
19. Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A. Evaluation of template-based models in CASP8 with standard measures. Proteins 2009;77 Suppl 9:18–28.
20. Keedy DA, Williams CJ, Headd JJ, Arendall WB, 3rd, Chen VB, Kapral GJ, Gillespie RA, Block JN, Zemla A, Richardson DC, Richardson JS. The other 90% of the protein: assessment beyond the Calphas for CASP8 template-based and high-accuracy models. Proteins 2009;77 Suppl 9:29–49.
21. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. Proteins 2007;69 Suppl 8:38–56.
22. Kryshtafovych A, Fidelis K, Moult J. CASP8 results in context of previous experiments. Proteins 2009;77 Suppl 9:217–228.
23. Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. Nature 1996;381:272.
24. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: A program to check the stereochemical quality of protein structures. J Appl Cryst 1993;26:283–291.
25. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB, 3rd, Snoeyink J, Richardson JS, Richardson DC. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res 2007;35:W375–383.
26. Shi S, Pei J, Sadreyev RI, Kinch LN, Majumdar I, Tong J, Cheng H, Kim BH, Grishin NV. Analysis of CASP8 targets, predictions and assessment methods. Database (Oxford) 2009;2009:bap003.
27. Shang Y, *et al.* MUFOLD-WQA: A new selective consensus method for Quality Assessmenmt in Protein Structure Prediction. Proteins 2011;79(Suppl 10):185–195.
28. Xu J, *et al.* RaptorX: Exploiting structure information for protrein alignment by statistical inference. Proteins 2011;79(Suppl 10):161–171.
29. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. Proteins 2003;53 Suppl 6:352–368.
30. Xu D, *et al.* A multi-layer approach for protein structure prediuction and model quality assessment. Proteins 2011;79(Suppl 10):172–184.
31. Hildebrand A, Remmert M, Biegert A, Soding J. Fast and accurate automatic structure prediction with HHpred. Proteins 2009;77 Suppl 9:128–132.
32. Zhang Y, *et al.* Automated protein structure modeling in CASP9 by I-TASSER pipeline. Proteins 2011;79(Suppl 10):147–160.
33. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21:951–960.
34. Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 2010;11:431.
35. Margelevicius M, Venclovas C. Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparison. BMC Bioinformatics 2010;11:89.
36. Scheeff ED, Bourne PE. Application of protein structure alignments to iterated hidden Markov model protocols for structure prediction. BMC Bioinformatics 2006;7:410.
37. Postma PW, Lengeler JW, Jacobson GR. Phosphoenolpyruvate:carbohydrate phosphotransferase systems of bacteria. Microbiol Rev 1993;57:543–594.
38. Krieger E, Joo K, Lee J, Raman S, Thompson J, Tyka M, Baker D, Karplus K. Improving physical realism, stereochemistry, and sidechain accuracy in homology modeling: Four approaches that performed well in CASP8. Proteins 2009;77 Suppl 9:114–122.
39. Kryshtafovych A, Fidelis K, Moult J. Progress from CASP6 to CASP7. Proteins 2007;69 Suppl 8:194–207.
40. Kryshtafovych A, Fidelis K, Moult J. CASP9 results compared to previous CASP experiments. Proteins 2011;79(Suppl 10):196–207.
41. McGuffin L, *et al.* Automated tertiary structure prediction with accurate local model quality assessment using the IntFOLD-TS methos. Proteins 2011;79(Suppl 10):137–146.
42. Poupon A, Janin J. Analysis and prediction of protein quaternary structure. Methods Mol Biol 2010;609:349–364.
43. Schmidt T, Haas J, Gallo Cassarino T, Schwede T. Assessment of ligand binding residue predictions in CASP9. Proteins 2011;79(Suppl 10):126–136.
44. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.
45. Biasini M, Mariani V, Haas J, Scheuber S, Schenk AD, Schwede T, Philippsen A. OpenStructure: a flexible software framework for computational structural biology. Bioinformatics 2010;26:2626–2628.
46. Wang G, Dunbrack RL, Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
47. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank : ensuring a single, uniform archive of PDB data. Nucleic Acids Res 2007;35:D301–303.
48. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol 2007;372:774–797.
49. Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, Leaver-Fay A, Baker D, Popovic Z, Players F. Predicting protein structures with a multiplayer online game. Nature 2010;466:756–760.