PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Assessment of template-based protein structure predictions in CASP10

Yuanpeng J. Huang,[1,2,3] Binchen Mao,[1,2,3] James M. Aramini,[1,2,3] and Gaetano T. Montelione[1,2,3]*

[1] Center for Advanced Biotechnology and Medicine and Department of Molecular Biology and Biochemistry, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854

[2] Department of Biochemistry and Molecular Biology, Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854

[3] Northeast Structural Genomics Consortium, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854

ABSTRACT

Template-based modeling (TBM) is a major component of the critical assessment of protein structure prediction (CASP). In CASP10, some 41,740 predicted models submitted by 150 predictor groups were assessed as TBM predictions. The accuracy of protein structure prediction was assessed by geometric comparison with experimental X-ray crystal and NMR structures using a composite score that included both global alignment metrics and distance-matrix–based metrics. These included GDT-HA and GDC-all global alignment scores, and the superimposition-independent LDDT distance-matrix–based score. In addition, a superimposition-independent RPF metric, similar to that described previously for comparing protein models against experimental NMR data, was used for comparing predicted protein structure models against experimental protein structures. To score well on all four of these metrics, models must feature accurate predictions of both backbone and side-chain conformations. Performance rankings were determined independently for server and the combined server plus human-curated predictor groups. Final rankings were made using paired head-to-head Student's *t*-test analysis of raw metric scores among the top 25 performing groups in each category.

Proteins 2014; 82(Suppl 2):43–56.

Key words: CASP10; protein structure prediction; GDT score; LDDT score; RPF DP scores; structural bioinformatics; homology modeling; comparative modeling.

INTRODUCTION

Template-based modeling (TBM) is an essential and highly successful approach for protein structure prediction. Recent advances, though generally incremental, provide both server and human curated methods with high reliability for protein structure prediction.[1] These TBM methods are having a high impact by providing accurate models useful in biological research.[2] The success of TBM has also been a primary driving force for structural genomics efforts aimed at structural coverage of domain families and biological pathways.[3–9] TBM is also a powerful technique for estimating phases in X-ray crystal structure determination by molecular replacement.[10–12] For these reasons, TBM forms an essential component of the critical assessment of protein structure prediction (CASP) experiment.

An important activity of the CASP program is the assessment of models and ranking of the performance by various predictor groups. A hallmark of this process is the involvement of independent assessors. A key feature of the CASP model assessment process is that it is done

Additional Supporting Information may be found in the online version of this article.

Abbreviations: AU, CASP10 assessment units, corresponding to regions of experimental structures used in assessing model predictions; FM, free modeling; TBM, template-based modeling; TBM hsAUs, all 112 TBM AUs, including AUs for which models were provided by either server or human-curated predictors; TBM soAUs, 55 AUs for which only predictions by servers were provided and assessed; TBM hAUs, the subset of 57 hsAUs for which human-curated predictions were provided and assessed.

Grant sponsor: Protein Structure Initiative of the National Institutes of Health; Grant number: U54-GM094597.

Y.J.H. and B.M. contributed equally to this work.

*Correspondence to: Gaetano T. Montelione, Center for Advanced Biotechnology and Medicine, Rutgers, The State University of New Jersey, 679 Hoes Lane, Piscataway, NJ 08854-5638.
E-mail: guy@cabm.rutgers.edu

Received 30 July 2013; Revised 10 November 2013; Accepted 19 November 2013
Published online 10 December 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24488

in a "blind" fashion, whereby assessors do not know the identity of each predictor group, which are identified only by a group number. The identities of the competing groups are provided to assessors only after the assessment process is completed.

For CASP10, participants were asked to provide full atomic models of predicted structures, including side-chain atoms. Accurate side-chain structures are often critical for using predicted models in biological applications. Following an emphasis established in CASP8[13] and CASP9,[1] the TBM assessment process of CASP10 used a combination of metrics which together assess the accuracy of both the backbone and side-chain structures of predicted models.

The assessment of free modeling (FM) protein structure predictions is presented in another article in this same special issue.[14] In this article, we describe the TBM assessment results of CASP10. Our analysis largely followed the protocols laid down in CASP8[13,15] and CASP9.[1,16] In particular, we followed the general procedure using standard measures for TBM assessment outlined by Cozzetto *et al.*,[15] and considerations of structure prediction accuracy metrics and statistical comparison tests outlined by Mariani *et al.*[1] for the CASP9 TBM assessment.

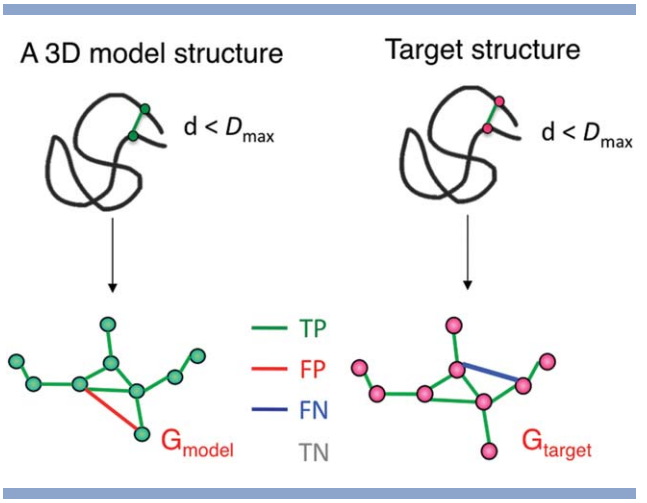

Figure 1

Comparison of models using distance networks. G_{target} is generated from the target structure and G_{model} is built from a prediction model. Edges that are present in both G_{target} and G_{model} are true positives (TP). Edges present in G_{target}, but not in G_{model} are false negatives (FN). Edges that are not present in both G_{target} and G_{model} are true negatives (TN). Edges present in G_{model}, but not in G_{target} are false positives (FP). From these, the recall R, precision P, and *F*-measure F are computed as described in the "Methods" section.

METHODS

Defining assessment units

Assessment units were defined based on careful manual analysis of the experimental structures and potential templates, as outlined in the accompanying article by Taylor *et al.*[17] Experimental NMR structures, as well as some X-ray crystal structures, were trimmed back to include only the consistently well-defined regions of the structure using the expanded FindCore algorithm, as described by Snyder *et al.*[18] (accompanying article).

Numerical automated structure quality assessment scores

All structure quality assessment scores were computed by the CASP Prediction Center.[19] GDT and GDC scores were computed for all predictions using the latest version of LGA.[19,20] The Prediction Center also provided large-scale calculations of LDDT,[1] RPF, Sphere Grinder,[19] MolProbity,[13,21,22] and Prosa[23] scores. LDDT scores, which compare the interatomic distance matrices between a predicted model and the experimental structure,[1,24] were computed with a distance cutoff of 15 Å, which is larger than the 5 Å cutoff used in CASP9.[1] They are referred to throughout this study as LDDT-15 scores. The "--rm=zeroocc" parameter was not used when computing the LDDT score. This means that atoms with zero occupancy were actually included in the LDDT score calculation. As there are not many predictions with zero occupancy atoms, this omission had minimal effect. Sphere Grinder used a 6 Å radius for the sphere, and a 2 Å rmsd cutoff. While these other metrics are described in other publications, the RPF metric for comparing predicted models against experimental structures is described in the next section.

RPF and DP scores

The RPF method was originally developed as a quality assessment tool for protein NMR structures. The NMR RPF method calculates the Recall, Precision, *F*-measure and discriminating power (DP) scores, by measuring the agreement between all proton–proton distances ≤ 5 Å, and the NMR NOESY and chemical shift data.[25] Specifically, the DP score measures how well a structural model fits with the NMR data (i.e., NOESY peak list and chemical shift data), normalized by the *F*-measure score that would be obtained by a random coil. A strong correlation has been observed between RPF scores of NMR structure models generated by automated NMR data analysis methods and GDT-TS/RMSD values relative to the corresponding crystal structures or manually refined NMR structures.[25–27]

The RPF method was adapted here as one of the assessment scores for template-based assessment in CASP10. Instead of measuring the agreement between structure models and NMR data, the modified RPF score used for CASP10 measures the agreement between the coordinates of a prediction model and an experimental (X-ray or NMR) structure of the AU.

RPF scores were computed using the following procedure: From the target protein structure, a network G_{target}

is built, as illustrated in Figure 1. Vertices (V) represent all N or C atoms from target structure and edges (E_{target}) connect the vertices if their corresponding distance in the model structure is $\leq D_{max}$. A similar distance network G_{model} is built from the prediction model. The agreement between the two structures is reduced to compare the differences between the two graphs G_{model} (derived from the prediction model) and G_{target} (derived from the target structure).

The RPF score for comparing predicted and experimental models is described for the first time in this article. TP, FN, FP, and TN are defined in Figure 1. RPF metrics are then calculated as: Recall (model, AU) = TP/(TP + FN), which measures the percentage of close distance atom pairs from the experimental AU structure that are also close in the prediction model; Precision (model, AU) = TP/(TP + FP), which measures the percentage of close atom pairs from the prediction model that are also close in the experimental AU structure; and F-measure (model, AU) = $[(1 + b)^2 \times$ Precision (model, AU) $\times$ Recall (model, AU)]/$[(b^2 \times$ Precision (model, AU) + Recall (model, AU)]. We use $b = 2$ to weight the Recall higher than Precision. Operationally, we also calculate a discriminating power (DP) score as DP = [F(model, AU) − F(random, AU)]/[1 − F(random, AU)], where F(random, AU) is the F-measure score calculated by comparing the distance networks of a random coil with the distance network of the AU structure. The distance network of a random coil is computed using atom distances expected for a freely rotating polypeptide chain model, as described by Flory and coworkers.[28] The Recall, Precision, and F-measure scores, as well as the normalized RPF score, the DP score, range from 0 to 1.0. The RPF score reported throughout this study is the DP score computed from the RPF algorithm.

Supporting Information Figure S1 shows that at short distance cutoffs (e.g., the 5 Å cutoff used for the NMR RPF assessment score), RPF is dominated by local side-chain packing information. As the distance cutoff increase, the correctness of fold starts to contribute to the RPF score. We choose distance cutoff of 9.0 Å, which seems to be a good balance of both global fold (main chain conformation) and also local side-chain and core side-chain packing information. For difficult targets with poor overall fold accuracy, the RPF score will be dominated by the main chain conformation, not the local side-chain packing. In comparing models with accurate folds, the RPF score will assign higher scores for models with better local side-chain conformations and core side-chain packing. Scores computed using a distance cutoff of 9 Å are referred to as RPF-9 scores.

If there are atoms missing from the predicted model but present in the experimental structure, interatomic distances for the missing pairs are set to 10,000 Å in the G_{model}, so as to count them as recall violations; that is, for interatomic distances $\leq D_{max}$ in the experimental structure, missing atoms in the predicted model were penalized in the RPF score by treating them as having distance $> D_{max}$ to all other atoms. In this way, the predicted model was penalized for the number of close distances $\leq D_{max}$ in the experimental structure that it fails to predict, including those it fails to predict because of missing atoms. This assessment metric thus encourages groups to submit complete atomic coordinates.

The RPF measure used in CASP10 was adapted from the NMR RPF measure developed for comparing protein models against protein NMR data, which do not use oxygen atoms for the practical reason that O atoms are not observed in the NMR data. In the course of developing of RPF for CASP10, we did include all oxygen atoms for comparison, and we found that it enhanced the weight for backbone atom positions, and decreased its sensitivity to side chain core packing. In our opinion, it was therefore preferable to exclude the oxygen atoms in the CASP10 RPF-9 scores.

Assessment of predicted protein models

Several metrics of protein structure accuracy were initially assessed. In keeping with the goal of encouraging predictors to submit high accuracy models, the standard GDT-TS assessment score of the LGA program,[20] with distance cutoffs of 1, 2, 4, and 8 Å, was replaced with the "high resolution" GDT-HA score,[15,29] with distance cutoffs of 0.5, 1.0, 2.0, and 4.0 Å. Besides GDT-HA, several other metrics were considered, including (i) GDC-all,[13] (ii) LDDT,[1] (iii) RPF, and (iv) Sphere Grinder (SphGr).[19]

The GDT-HA and GDC-all scores are global measures of the agreement between a predicted model and the experimental structure, with GDT-HA reflecting the accuracy in placing Cα positions, and GDC-all including information about the positions of side-chain carbon atoms. On the other hand, the LDDT, RPF, and SphGr metrics are more sensitive to local structure accuracy and core packing. These intuitive perspectives were confirmed by examining many examples of superimposed models (and model fragments) with experimental structures, and comparing the corresponding scores.

The knowledge-based metrics such as ProsaII[23] and MolProbity[21,22] are valuable for assessing the physical reasonableness of molecular models in the absence of a "gold standard" by which to assess structural accuracy. However, it is well known that incorrect structures can have good ProsaII or MolProbity scores; for example, a perfect alpha helix prediction can have excellent MolProbity scores even if the true structure is a beta strand. In the CASP TBM assessment, high-accuracy experimental X-ray crystal or NMR structures are available. Geometric comparisons between predicted models and these experimental structures provide a more rigorous basis for assessing the accuracy of a predicted model than knowledge-

based metrics, particularly when using structural similarity metrics like GDC-all, LDDT, and RPF which include both backbone and side-chain atom positions in comparisons between the predicted model and the experimental structure. Hence, it was decided to carry out numerical assessment for ranking predicted models and predictors using only metrics that compare atomic coordinates of models with the corresponding experimental models; that is, measures of the structural accuracy. Notwithstanding this decision, knowledge-based MolProbity scores were also computed for each predicted model and are included in the summary of assessment metrics.

Based on the considerations outlined in the preceding two paragraphs and our preliminary analyses, the numerical automated assessment of models was done using an equal weighting of 4 scores: two global alignment scores GDT-HA and GDC-all, and two superimposition-independent locally oriented scores, RPF and LDDT, that are highly sensitive to side-chain atom positions and to the accuracy of local structure and side-chain packing. The LDDT metric was used with cutoff of 15 Å (LDDT-15), and the RPF metric was used with cutoff of 9 Å (RPF-9), as explained in the Supporting Information section. Our initial analyses also include SphereGrinder scores, but as omitting these scores did not have a significant impact on the rankings, our final analysis included only the four scores described above. Overall, the resulting CASP10 TBM composite assessment score is therefore quite similar to that used in CASP9,[1] which included an equal weighting of GDT-HA, GDC-all, and LDDT-5 (5 Å cut off) scores. However, the addition of the RPF-9 score to the assessment provided more weight in the total score on local side-chain conformation and core side-chain packing relative to overall backbone structure.

Selection of models for assessment

As in CASP8 and CASP9, for each AU predicted by each group, only the model designated as "Model 1" (or the one with the lowest index if no "Model 1" was available) was used for automated numerical assessment. Models containing < 20 residues were excluded from assessment. For predictions that included multiple fragments, the fragment with the longest overlap with the target AU was used.

RESULTS AND DISCUSSION

As CASP has evolved there has been increasing emphasis on the accuracy of predicting realistic protein structure models.[1,13,29,30] The goal is to encourage CASP predictions on the TBM class of targets towards the same standards used for experimental protein structure determinations, *including accurate positions for all heavy (C, N, O, and S) atoms*. In CASP10 this philosophy was continued, and predictors were instructed that evaluation would include not only an assessment of the accuracy of backbone Cα atom positions, but also the accuracy of side-chain heavy atom (i.e., nonproton) positions.

Summary of targets, predictions, and assessment units

During CASP10, 114 protein sequences were released by the Prediction Center. Seventeen of these were later cancelled for various reasons (as outlined in Taylor *et al.*[17]), leaving 97 proteins and protein domains which ranged in size from 33 residues up to 770 residues. Following the process used in CASP7 through CASP9,[31–33] targets were first split into assessment units (AUs) and assigned to either the "FM" or "TBM" assessment groups, as outlined in the accompanying article by Taylor *et al.*[17] The TBM AUs ranged in size from 24 residues (T0709-D1) to 498 residues (T0645-D1 and T0664-D1). In total, 112 AUs associated with 91 experimental protein structures were assigned to the TBM group.

In CASP10, 41,740 predictions were submitted for the 112 assessed TBM AUs, by a total of 150 predictor groups. Of these groups, 69 were registered as prediction servers and 81 were expert "human-curated" predictor groups. The CASP process allows each submitter to provide up to five alternative models. However, the primary TBM assessment considered only the single model designated as Model 1; this is supposedly the best model. Hence while more than 40,000 predicted models were submitted, only 10,287 predicted models for the 112 AUs were considered in the numerical rankings. As outlined below, studies were also carried out of the impact of model selection on rankings. These hypothetical rankings required consideration of accuracy scores for all 41,740 predicted models.

As in CASP8 and CASP9, protein sequences for which templates could be easily identified by sequence-based methods were classified as "server-only targets" at the time the target was released by the CASP Prediction Center. The goal of this designation is to allow human expert groups to focus their attention on the remaining more challenging TBM and FM targets.[34] If a target was released as "server-only," then all AUs from this target were defined as soAUs. If a target was released as "all-group" (or human/server, hs), then all AUs from this target were hsAUs. Accordingly, 55 AUs were designated as server-only TBM AUs (soAUs). Server groups (S groups) were assessed on prediction of all 112 TBM AUs (i.e., the human *or* server hsAUs), while human expert groups (H groups) were assessed only on predictions for the subset of 57 TBM AUs (hAUs), excluding the 55 soAUs. Following the presentation of results for CASP8 and CASP9, the results of assessing the 112 hsAUs by the S groups, and the subset of 57 hAUs by the H and S groups, are

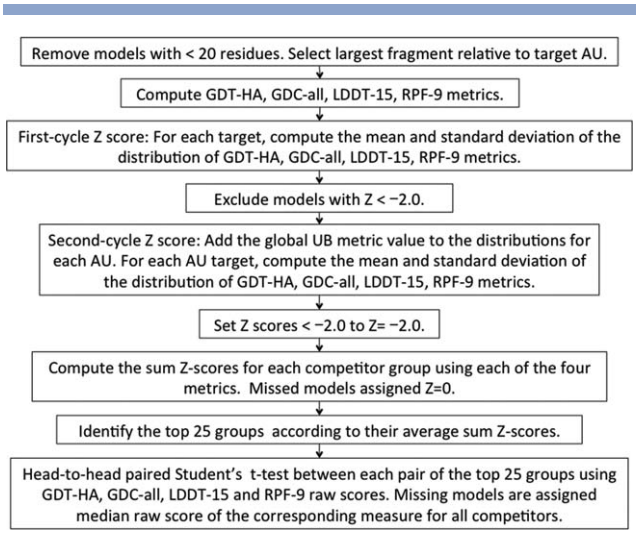

Figure 2

Flowchart of the procedure used for CASP10 TBM numerical assessment and ranking of predictor groups.

presented separately. However, the same metrics and methods, outlined in the following section, were used to assess all predictions.

Automated numerical assessment

For every submitted model, the CASP10 Prediction Center[19] computed raw scores for each metric and AU. These scores, for every model submitted, can be found on the CASP10 Prediction Center web site (http://www.predictioncenter.org/casp10/results.cgi). Where submitted models spanned more than one AU, numerical scores were computed relative to each AU.

The automated numerical assessment of predictions and ranking of predictor groups was carried out following the same general strategy used in previous recent CASP comparative modeling or TBM assessments.[1,13,15,30,35] A flow chart of the process is presented in Figure 2.

For each submitted model and corresponding AU, raw GDT-HA, GDC-all, LDDT-15, and RPF-9 scores were compiled. Next, for each AU the mean and standard deviations of these scores were computed. The mean and standard deviations for each of these distributions (for each of 112 hsAU's and each of the four metrics) were then used to assign a Z score for each metric to each prediction model.

Ranking based on Z scores

As in CASP8 and CASP9, these initial Z scores were then used to eliminate the most inaccurate models for the Z score analysis. The motivation for doing this is to encourage predictors to explore new methods, and to minimize penalties that incur due to bad models that may result from the exploration of new methods.[36] In recent TBM CASP assessments,[1,15] the resulting models were then used to recalculate Z scores. As an additional motivation to minimize penalties due to poor models, models with Z scores < 0 were assigned $Z = 0$, and the resulting Z scores were used to compute a composite Z score for the predictor group, accounting for the performance on all models with all metrics. This composite score was then used to rank the predictor group.

As in CASP8 and CASP9, models with Z scores < -2 in first cycle of Z score analysis were excluded from the second-cycle Z score analysis (Fig. 2). However, for CASP10, final Z scores after the second cycle were computed somewhat differently. The intention of assigning $Z = 0$ to Z scores < 0 was to encourage CASP participants to submit predictions for difficult targets. Binning the targets into groups based on their difficulty, as measured by the maximum GDT-TS value obtained by any group for that target, we observed a high frequency of negative Z scores for easier targets. For these CASP10 targets, assigning $Z = 0$ to Z scores < 0 had the unintended consequence of improving scores for groups than made relatively poor predictions *on easy targets*; that is, AU targets for which other groups submitted accurate predictions. Interestingly, when some top scoring groups did poorly, they tended to do poorly on these easier targets; the process of assigning $Z = 0$ to Z scores < 0 tended to lessen the impact of these poorer models which in fact should contribute negatively to the ranking. For this CASP10 assessment, this effect was ameliorated by instead setting models with Z scores < -2 after the second cycle to $Z = -2$.

Our analysis of the Z score distributions also revealed another caveat. In the CASP8 analysis, it was pointed out that for some AUs for which no groups provided good models, the Z score can lead to an overestimate of performance.[15] For the difficult AUs with best GDT-TS scores $\leq 50\%$, for which most predictions were quite poor, a less-poor prediction often resulted in a significantly positive Z score; that is, the prediction was significantly better than most, but was still a very poor structure. For example, a significantly better-than-average GDT-HA score for a poor model might result in a significantly better Z score. This effect could be addressed by including in each distribution for each metric an "ideal score"; that is, the best score (UB) obtained for the corresponding metric for any model, on any AU, by any predictor. Figure 3 demonstrates how including a global UB raw score in the distribution used to compute Z scores for each metric suppresses high Z scores otherwise obtained for poorly modeled difficult AUs. In the examples shown in Figure 3(A,B), where a poor model prediction was substantially better than even less accurate predictions, inclusion of the UB score in the distributions suppressed the high estimate of performance by the Z score measure. Data summarized in Figure 3(C) demonstrate how the inclusion of UB scores reduces the high

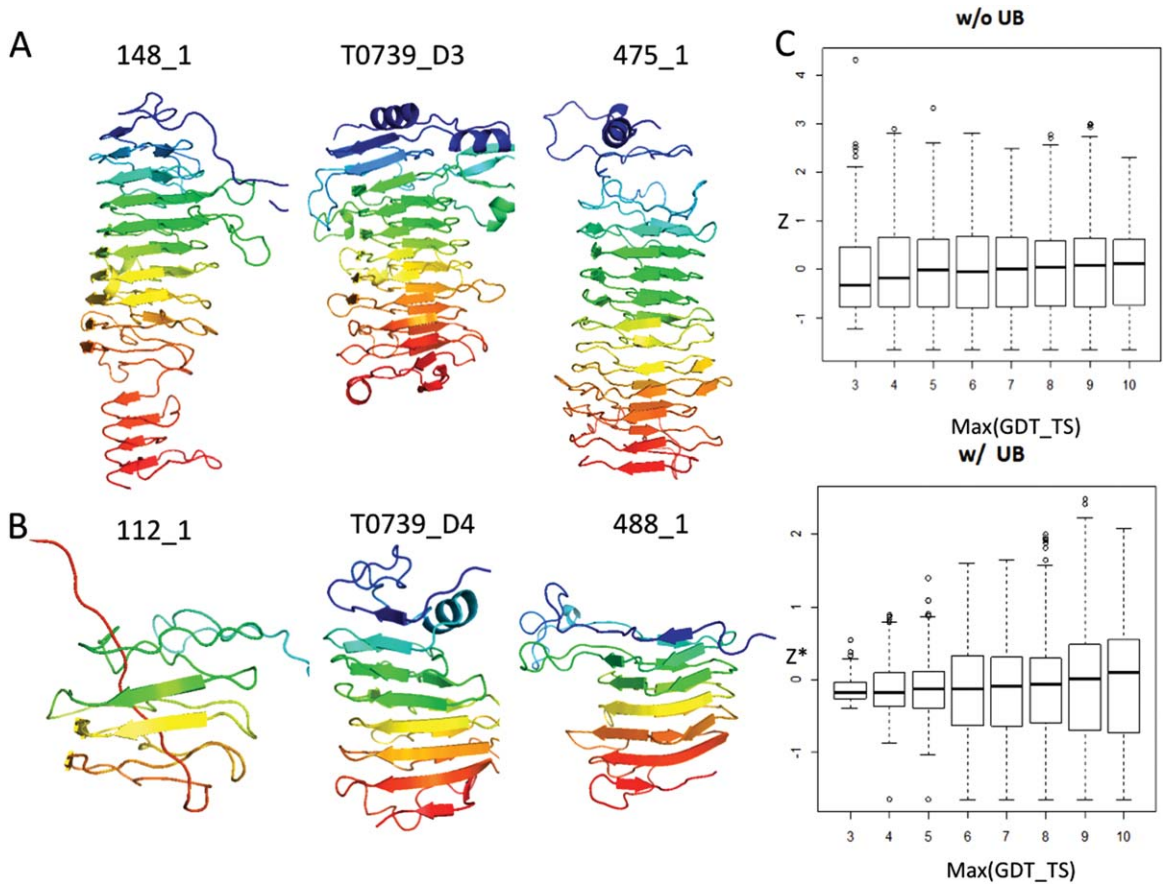

Figure 3

Impact on Z scores of including the global maximum raw score UB. In the example, AUs shown in Panels A and B, prediction models with low GDT-HA scores have high GDT-HA Z scores, even though all predictions for the corresponding AU are poor. Inclusion of the best GDT-HA score obtained for any AU (UB) in the raw score distribution, resulting in a modified Z score, Z^*, suppresses the high Z scores for these poor models. **A**: Center—the experimental structure of AU T0739_D3. Left—model 148_1(GDT-HA = 11.2/Z = 1.99/Z^* = 0.513). Right—model 475_1 (GDT-HA = 9.07/Z = 1.18/Z^* = 0.268). **B**: Center—the experimental structure of T0739_D4. Left—model 112_1(GDT-HA = 13.8/Z = 4.56/ Z^* = 0.811). Right—model 488_1 (GDT-HA = 10.7/Z = 2.71/Z^* = 0.445). **C**: Top—the Z score distributions calculated without using the UB, binned into classes of AUs based on the maximum GDT-TS (max GDT-TS score) obtained by any predictor group. For example, the left-most bin (labeled 3) presents box plots of Z scores for the AU class with max GDT-TS < 30%. Bottom—the modified Z^* score distribution which includes the UB raw score in the Z score analysis of each AU. Using a UB raw score in the distribution of scores for each AU significantly lowers the high Z scores otherwise obtained for poor predictions of the more difficult AUs.

Z scores obtained for predictions of AUs for which the best GDT-TS scores (i.e., Max GDT-TS) were $\leq 50\%$, with less effect on Z scores of AUs for which the best GDT-TS scores were $> 50\%$.

With the adjustments outlined above, including a UB "best raw score" in each distribution, recomputing Z scores, and then setting $Z = -2$ for $Z < -2$, we computed Z scores for each "Model 1" of each predictor group, for each of the four metrics (GDT-HA, GDC-all, LDDT-15, and RPF-9). The resulting Z scores for each metric and predictor group were then summed. Missing models were assigned a Z score of 0, generously minimizing the penalty for not submitting any prediction for an AU. These results are summarized for the 112 hsAUs in Supporting Information Table S-I, and for the 57 hAUs in Supporting Information Table S-II. The top 25 performing groups are listed in Table I for the server predictors and in Table II for the human curated predictors.

The sum of model Z scores for each metric by each group (Tables I and II) provides a useful assessment of performance with respect to that particular metric. However, for ranking the performances of the predictor groups it is useful to have a single composite performance score. For CASP10, the composite performance score (Sums in Tables I and II) was computed by averaging, for each predictor group, the summed Z scores for the four metrics.

An average performance score per AU was then calculated by dividing this composite score by the number of AUs. This average performance score for each group was computed in two ways: (i) dividing the composite performance score by the total number of available AUs (Avg-a) (i.e., 112 or 57), or (ii) dividing by the total

Table I
Sum and Average Z-scores for Top 25 Performing Server Predictor Groups – 112 hsAUs

Group	Name	N_model	GDT-HA	GDC-all	RPF	LDDT	Sum	Avg-a	Avg-s	MolPro
330s	BAKER-ROSETTASERVER	112	52.03	62.49	77.60	75.18	66.83	0.60	0.60	219.08
035s	Zhang-Server	112	54.21	48.41	78.52	66.59	61.93	0.55	0.55	15.62
108s	PMS	112	37.21	49.23	72.74	74.08	58.32	0.52	0.52	2.03
114s	QUARK	111	50.68	41.91	70.40	60.60	55.90	0.50	0.50	9.40
370s	HHpred-thread	111	44.39	48.89	50.60	58.05	50.48	0.45	0.46	−144.25
122s	RaptorX-ZY	112	43.44	42.14	53.96	47.22	46.69	0.42	0.42	−70.08
430s	HHpredA	112	44.20	46.70	45.30	49.21	46.35	0.41	0.41	−137.90
223s	HHpredAQ	112	40.82	43.87	44.73	48.88	44.58	0.40	0.40	−139.75
486s	RaptorX	112	42.91	43.50	42.62	36.23	41.32	0.37	0.37	−36.43
424s	MULTICOM-NOVEL	112	31.30	34.88	45.11	47.34	39.66	0.35	0.35	27.56
125s	MULTICOM-REFINE	112	27.70	33.03	39.74	42.52	35.75	0.32	0.32	30.97
081s	MULTICOM-CLUSTER	112	24.39	30.72	39.49	41.41	34.00	0.30	0.30	30.46
335s	TASSER-VMT	112	29.13	24.93	48.74	28.75	32.89	0.29	0.29	−88.38
103s	PconsM	112	24.21	25.59	43.61	37.44	32.71	0.29	0.29	27.14
488s	chunk-TASSER	112	24.85	29.28	38.73	36.42	32.32	0.29	0.29	−28.19
292s	Pcons-net	112	13.93	20.25	32.86	28.66	23.93	0.21	0.21	65.23
286s	Mufold-MD	112	10.72	16.35	32.47	30.63	22.54	0.20	0.20	93.95
222s	MULTICOM-CONSTRUCT	112	14.40	19.42	26.07	29.11	22.25	0.20	0.20	16.48
333s	MUFOLD-Server	112	16.05	15.96	27.70	23.79	20.88	0.19	0.19	−13.69
261s	Seok-server	112	13.31	23.50	12.71	25.40	18.73	0.17	0.17	75.22
411s	FALCON-TOPO	112	6.67	7.67	17.54	11.72	10.90	0.10	0.10	−30.77
456s	FALCON-TOPO-X	112	3.35	5.14	11.72	5.90	6.53	0.06	0.06	−36.42
124s	PconsD	111	0.89	−0.42	14.28	8.15	5.73	0.05	0.05	0.36
348s	Phyre2_A	112	7.88	5.33	2.20	6.29	5.43	0.05	0.05	−82.57
413s	ZHOU-SPARKS-X	112	−3.20	−5.76	11.41	8.48	2.73	0.02	0.02	−68.11

The columns labeled GDT-HA, GDC-all, RPF, and LDDT are the sum of Z scores across all models submitted by each predictor group. The Sum column is the average of the sum of Z scores for the four metrics assessed. The Avg-s and Avg-a scores are the Sum scores divided by the number of AUs for which a model was submitted (s) by each predictor group, and the total number of AUs used for assessment (a), respectively. These scores are identical for predictor groups who submitted models for all 112 hsAUs.

Table II
Sum and Average Z-scores for Top 25 Performing Predictor Groups −57 hAUs.

Group	Name	N_model	GDT-HA	GDC-all	RPF	LDDT	Sum	Avg-a	Avg-s	MolPro
237	zhang	57	37.64	33.50	50.23	46.15	41.88	0.74	0.74	1.63
27	LEEcon	57	35.60	36.14	45.39	47.43	41.14	0.72	0.72	12.84
035s	Zhang-Server	57	32.85	31.00	44.68	40.19	37.18	0.65	0.65	−2.33
130	Pcomb	57	29.92	28.99	44.96	41.68	36.39	0.64	0.64	22.33
197	Mufold	56	30.46	29.74	42.41	39.20	35.45	0.62	0.63	−9.75
79	TASSER	57	34.34	32.49	40.59	33.92	35.34	0.62	0.62	−32.59
267	Pcons	56	28.76	28.78	42.48	40.73	35.19	0.62	0.63	15.26
489	MULTICOM	57	29.85	29.53	40.17	39.02	34.64	0.61	0.61	9.01
344	Jones-UCL	56	26.78	28.72	43.08	37.82	34.10	0.60	0.61	−50.10
114s	QUARK	56	29.68	25.98	38.50	35.87	32.51	0.57	0.58	−5.65
301	LEE	57	27.53	29.38	34.98	37.52	32.35	0.57	0.57	6.37
477	BAKER	57	31.46	29.24	35.34	33.04	32.27	0.57	0.57	104.97
475	CNIO	57	27.86	23.52	39.78	37.69	32.21	0.57	0.57	−2.34
350	Kloczkowski_Lab	57	26.74	23.82	40.01	36.94	31.88	0.56	0.56	18.79
490	Zhang_Refinement	57	26.03	24.81	33.25	32.61	29.18	0.51	0.51	16.67
294	chuo-repack	57	23.19	20.55	36.79	34.19	28.68	0.50	0.50	−6.90
458	Sternberg	57	25.94	23.30	32.25	31.66	28.29	0.50	0.50	−10.66
365	chuo-fams	57	22.50	21.08	35.92	31.92	27.86	0.49	0.49	4.58
428	PconsQ	56	22.08	20.09	33.92	33.43	27.38	0.48	0.49	7.49
434	chuo-fams-consensus	57	20.19	15.16	32.68	31.41	24.86	0.44	0.44	−7.49
481	Chicken_George	57	21.70	18.50	31.02	28.17	24.85	0.44	0.44	5.58
122s	RaptorX-ZY	57	25.88	24.34	24.68	21.98	24.22	0.43	0.43	−44.38
45	Zhang_Ab_Initio	57	21.07	20.37	28.28	26.96	24.17	0.42	0.42	4.98
285	McGuffin	55	18.28	16.26	30.52	27.92	23.25	0.41	0.42	−1.60
405	Mufold2	54	18.98	18.36	28.92	25.93	23.05	0.40	0.43	−17.03

The columns labeled GDT-HA, GDC-all, RPF, and LDDT are the sum of Z scores across all models submitted by each predictor group. The Sum column is the average of the sum of Z scores for the four metrics assessed. The Avg-s and Avg-a scores are the Sum scores divided by the number of AUs for which a model was submitted (s) by each predictor group, and the total number of AUs used for assessment (a), respectively. These scores are identical for predictor groups who submitted models for all 57 hAUs.

number of AUs submitted by the group (Avg-s). The former method penalizes groups that did not submit predictions for all AUs. For the top performing groups, all of which provided predictions for 111–112 hsAUs (Table I) [or for 56–57 of the hAUs (Table II)], these average values were approximately the same. The rankings shown in Tables I and II are based on the sum average Avg-a, which slightly penalizes the groups that did not submit predictions for all of the available AUs.

The sum MolProbity[21,22] Z scores for each predictor group is also tabulated for the top 25 performing groups in Tables I and II (and in Supporting Information Tables I and II for all groups). Although these MolProbity scores were not used in ranking predictor groups, they provide a knowledge-based assessment of structure quality. In particular, the Baker group consistently provided models with good MolProbity scores.

Head-to-head paired student's t-test

The Z score ranking was not used to provide a final ranking of predictor groups. As in CASP9,[1] the Z score ranking (based on the sum average Avg-a scores) was only used to identify the top 25 groups, which were then selected for a more detailed assessment accounting for the statistical significance of ranking one group above another. The predictions of these 25 groups were then compared in a direct head-to-head statistical analysis on common targets. In this analysis, each score distribution (GDT-HA, GDC-all, LDDT-15, RPF-9) was considered separately. Using each of these metrics independently, the raw score distributions for each of the AUs from each predictor group were compared with the corresponding score distributions from other groups in the top 25 list, using the paired Student's t-test. The number of comparisons for which there was a statistically significant difference was then summed for each metric. For each score metric, each of the top 25 groups is compared with 24 other groups. Group A was assigned 1 point if its distribution of raw scores for that metric is significantly better ($P < 0.05$) than that of group B. Hence, the maximum score per metric is 24; and the maximum summed score for four metrics assessed is 96. For models with Z scores < -2, the raw score was set to the raw score value for that AU corresponding to $Z = -2$ (based on the second cycle Z score analyses summarized in Tables I and II). This consideration was relatively insignificant for most of the 25 top-performing groups, for which even the worst models generally have $Z > -2$. When a model was miss ing for a group, the median score for that measure was used as the raw score. This consideration was also insignificant, as most of the 25 top-performing groups submitted prediction models for all AU targets. These summed Student's t-test head-to-head pairwise scores (H2H Scores), plotted in Figure 4(A) (for the 112 hsAUs) and Figure 4(B) (for the 57 hAUs), provided the basis for determining the final ranking of predictor groups.

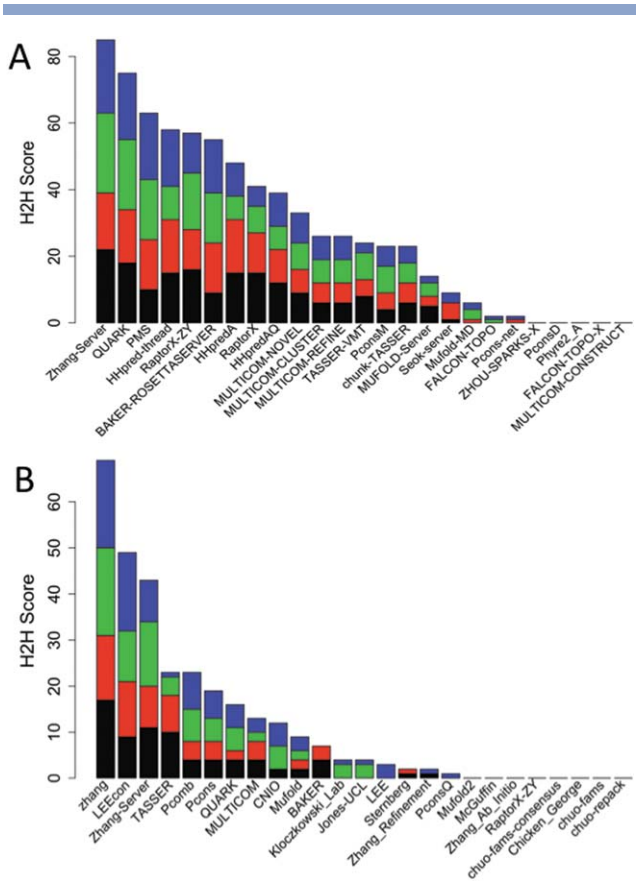

Figure 4

Ranking of top 25 server-only and human/server predictor groups. **A**: Head-to-head pairwise Student's t-test analysis on raw scores between 25 top-ranking server predictor groups for 112 hsAUs. **B**: Head-to-head pairwise Student's t-test analysis on raw scores between 25 top-ranking server/human predictor groups for 57 hAUs. Black, GDT-HA; red, GDC-all; green, RPF-9; blue, LDDT-15. Top-ranking 25 groups were identified based on average Z score Avg-a (Tables I and II).

It is important to recognize that the final ranking among the top 25 performing groups did not depend on the details of our methods for computing Z scores; the Z scores were used only to provide an overall ranking of all groups (presented in Supporting Information Tables I and II), and to identify the top 25 performing groups. The ranking within these top 25 groups was based on the raw scores for each of the four metrics, using paired Student's t-test in head-to-head comparisons of the predictions made by each group on the accuracy of the predictions of the same target by all of the other 24 groups. In all of the H2H analyses presented in this article, the order of the groups shown in the H2H analysis plots for which the distribution of raw prediction scores for each of the four metrics assessed were not statistically better than any of the other top 24 groups, are random. The relative rankings of these groups are determined by the

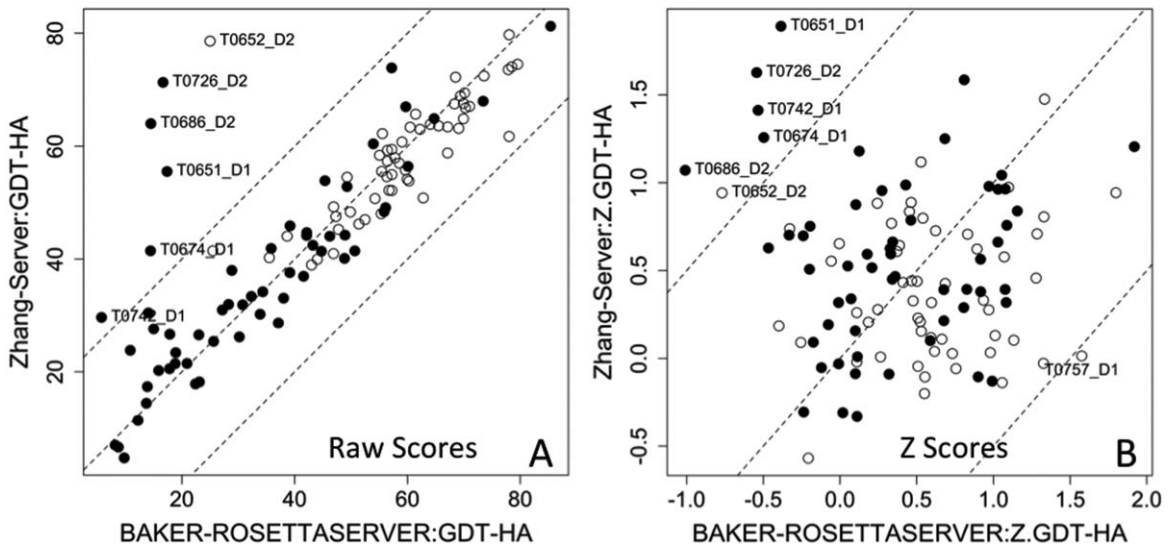

Figure 5

Raw metric scores provide more sensitivity to poor performance on more challenging hAUs. Comparison of performance by top-ranking groups shows how (**A**) GDT-HA raw scores are more sensitive to poor performance on more challenging hAUs than (**B**) GDT-HA *Z* scores. Close circles, hAU targets. Open circles, soAU targets. The three dashed lines of Panel A are at $y = x$ and $y = x \pm 20$ units, and the three dashed lines of Panel B are at $y = x$ and $y = x \pm 1.5$ units. Panel B is expanded to exclude regions of the plot with no data points.

Z score statistics of Supporting Information Tables S-I, S-II, and S-III.

H2H ranking—servers

Figure 4(A) summarizes the results of the head-to-head (H2H) Student's *t*-test analysis for the top 25 performing server groups, assessed on 112 AUs. The five top performing groups were Zhang Server (035s), QUARK (114s), PMS (108s), HHpred-thread (370s), and RaptorX-ZY (122s). These were followed by BAKER-ROSETTASERVER (330s), HHPredA (430s), RaptorX (486s), HHPredAQ (223s), and MULTICOM-NOVEL (424s).

Generally, the H2H ranking of Figure 4(A) correlates with the *Z* score ranking of Table I: for example, 035s > 108s > 114s > 370s > 122s > 430s > 223s > 486s > 424s. One notable outlier in this correlation is the BAKER-ROSETTASERVER (330s), which was top-ranked in the *Z* score analysis (Table I), but sixth-ranked in the H2H ranking [Fig. 4(A)]. Figure 5 compares GDT-HA raw score and *Z* score distributions for predictions made by Zhang Server (035s) and BAKER-ROSETTASERVER (330s). This analysis demonstrates that, based on raw GDT-HA scores [Fig. 5(A)], 035s had better performance on the generally more challenging hAUs (i.e., targets with low sequence identity with templates) than 330s. Performance by 330s was marginally better on soAUs, with easily identified templates of known structure. These results suggest that 035s was able to do a superior job than 330s in identifying distant templates. Although this same better performance on more challenging hAUs is evident in the *Z* score analysis [Fig. 5(B)], these *Z* scores are much less discriminating than the raw scores, particularly for the challenging hAUs with very low sequence identity with templates. The classification of soAUs and hAUs was made by the CASP Prediction Center based on the HHSearch probability score with potential templates, and may not always accurately reflect the actual difficulty of the targets. In addition, we observed many examples where *Z* scores overestimate prediction performance when the spreads of the metric for some specific targets are narrow (Supporting Information Figure S6). Overall, these results demonstrate the value of using the raw accuracy metric scores, through the H2H paired Student's *t*-test, in determining the final ranking.

We also explored the sensitivity of the H2H ranking against some extremely poor predictions. One approach for suppressing the high impact of those extreme outliers is to replace the Student's *t*-test with the nonparametric Wilcoxon-signed rank test. This statistic compares the rank, rather than value, of the raw score differences between two distributions, and is less sensitive to raw score outliers. As illustrated in Supporting Information Figure S2, suppression of outliers using the paired Wilcoxon rank sum analysis alters the ranking among the top 25 performing groups; for example, in this analysis BAKER-ROSETTASERVER (330s) becomes the top ranked server.

H2H ranking—human and/or servers

Figure 4(B) summarizes the results of the head-to-head (H2H) paired Student's *t*-test analysis for the top 25 performing human and/or server groups, assessed on the 57 hAUs. The score for the top-performing group, Zhang (237), was significantly higher than any other group. The next best performing groups were Leecon (027) and Zhang-Server (035s), followed by TASSER (079), Pcomb (130), Pcons (267), QUARK (114s), MULTICOM (489), CNIO (475), and Mufold (197).

For the Human and/or Server analysis, the H2H ranking was generally well-correlated with the *Z* score ranking (Table II); 237 > 027 > 035s > 130 > 197 > 79 > 267 > 489 > 344 > 114s. The differences between H2H and *Z* score ranking again could be traced to the higher sensitivity of raw score comparisons over *Z* score comparison, especially for hAUs with low sequence identity with templates.

H2H ranking—most difficult TBMs

We also ranked the human and/or server predictors using only the 15 "most difficult" AUs. Figure 6 is a bar plot of the maximum GDT-TS (Max GDT-TS) score obtained for each of the 112 hsAU's by any of the predictors. AUs with Max GDT-TS $\leq$ 50% are referred to here as TBM_hard AUs. Based on this analysis, 15 CASP10 AUs are defined as the "TBM_hard AUs": T0649_D1, T0653_D1, T0668_D1, T0671_D2, T0676_D1, T0678_D1, T0684_D1, T0690_D1, T0705_D2, T0717_D2, T0726_D1, T0732_D2, T0735_D1, T0739_D3, T0739_D4. Using only these 15 targets, we repeated the *Z* score analysis, shown in Table III for the top-ranked 25 groups, and in Supporting Information Table S-III for all groups. We also carried out head-to-head paired Student's *t*-test analysis of raw scores among the top-performing 25 groups. Figure 7 summarizes the results of the H2H, assessed on only the 15 TBM_hard AUs. For this set of targets, the top-performing groups were Zhang (237) and MULTICOM (489), with nearly identical performance that was significantly better than that for the other 23 groups. These were followed by MuFold (197), TASSER (079), Pcons (267), chuo-fams (365), RaptorX-ZY (122s), Pcomb (130), Kloczkowski (350), and Leecon (027). Significantly, of the 69 server predictors participating in CASP10, only one server, RaptorX-ZY (122s), performed well with these TBM_hard AUs.

It was particularly important to use multiple metrics of model accuracy in assessing performance on the TBM_hard AUs, as some groups scored well with some metrics, other groups scored well with different metrics. Only the two top-ranked groups Zhang (237) and MULTICOM (489) consistently scored well with all four metrics. This is distinct from the results with 112 hsAUs or 57 hAUs, in which there was generally consistent performance on all four metrics by the top scoring groups.

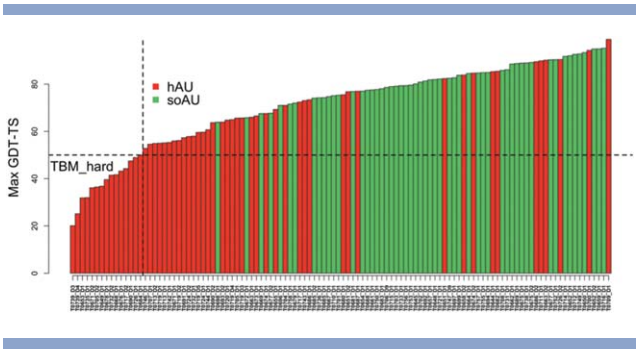

Figure 6

Classifying TBM_hard AUs. Bar plot of the maximum GDT-TS (Max GDT-TS) score obtained for each of the 112 AU's (indicated by AU target id) by any of the predictors. The 15 AU targets with Max GDT-TS $\leq$ 50% were identified as TBM_hard AUs.

Impact of distance-matrix methods in assessing low accuracy structures

Figure 7 shows that the superimposition-independent LDDT-15 and RPF-9 scores are helpful in distinguishing performance of predictor groups on the TBM_hard AU's. Also relevant was the complementary value of LDDT-15 and RPF-9 scores in ranking TBM_hard AUs; some predictor groups did best with LDDT-15 scores, while others did best with RFP-9 scores.

Unlike other metrics used in CASP10, the RPF-9 score is normalized against a freely rotating chain model, based on Flory polymer chain statistics.[25,28] For this reason, RPF-9 DP scores used in CASP10 are very discriminative against random structures. Random-like incorrect prediction models (i.e., low quality structures) will have RPF-9 scores very close to zero. Structures with incorrect secondary structures or incorrect folds can even have negative RPF-9 scores, indicating that they are even worse than random structures. Examples discussed in the Supporting Information demonstrate that the RPF-9 scores normalized to random structure (i.e., the RPF-9 DP scores) have stronger discriminating power than the GDT-HA scores against structures with random-like incorrect folds (Supporting Information Figure S3). Normalization against a set of well-defined decoy models, representing a random distribution of structures, has also recently been introduced into the LDDT score.[24] Additional examples comparing the sensitivity of global alignment (GDT and GDC-all) scores and superimposition-independent (LDDT and RPF) scores for assessing surface loop and interhelical packing inaccuracies are also presented in the Supporting Information Figure S4.

Figure 7 also shows that TASSER (079), RaptorX-ZY (122s), and MULTICOM (489) had the best performance on TBM_hard AUs with the GDC-all score, although the first two of these groups did not do so well with the

Table III
Sum and Average Z-scores for Top 25 Performing Predictor Groups −15 TBM_Hard AUs.

Group	Name	N_model	GDT-HA	GDC-all	RPF	LDDT	Sum	Avg-a	Avg-s	MolPro
489	MULTICOM	15	7.24	6.58	14.76	11.99	10.14	0.68	0.68	1.48
237	zhang	15	6.84	4.95	12.46	10.98	8.81	0.59	0.59	−0.84
197	Mufold	14	6.02	3.92	12.09	10.40	8.11	0.54	0.58	−6.17
267	Pcons	15	5.28	4.39	11.37	10.37	7.85	0.52	0.52	1.74
428	PconsQ	15	4.91	4.04	10.90	9.80	7.41	0.49	0.49	−0.36
365	chuo-fams	15	4.68	2.88	11.77	10.17	7.38	0.49	0.49	1.99
27	LEEcon	15	5.08	3.38	10.55	9.44	7.11	0.47	0.47	3.17
130	Pcomb	15	4.55	3.05	11.10	9.17	6.97	0.46	0.46	8.60
035s	Zhang-Server	15	5.18	3.55	9.85	8.88	6.87	0.46	0.46	−3.49
344	Jones-UCL	14	4.30	3.47	10.25	8.54	6.64	0.44	0.47	−13.90
294	chuo-repack	15	3.43	2.44	10.62	9.40	6.47	0.43	0.43	−3.02
350	Kloczkowski_Lab	15	3.18	2.37	10.84	9.00	6.35	0.42	0.42	−1.04
285	McGuffin	14	4.13	3.15	9.13	7.77	6.05	0.40	0.43	−0.45
475	CNIO	15	3.55	1.79	9.99	8.62	5.99	0.40	0.40	−1.58
79	TASSER	15	5.56	6.94	7.17	3.98	5.91	0.39	0.39	−2.92
434	chuo-fams-consensus	15	2.83	1.55	9.51	8.67	5.64	0.38	0.38	−0.49
114s	QUARK	14	3.83	2.73	8.45	7.31	5.58	0.37	0.40	−5.13
315	keasar	15	3.88	2.24	9.02	6.39	5.38	0.36	0.36	−15.88
122s	RaptorX-ZY	15	5.88	5.62	5.56	3.75	5.20	0.35	0.35	−14.48
45	Zhang_Ab_Initio	15	2.96	3.37	7.68	6.71	5.18	0.35	0.35	1.19
490	Zhang_Refinement	15	4.00	2.56	7.52	6.40	5.12	0.34	0.34	4.24
26	ProQ2clust	14	2.78	3.41	6.79	6.68	4.92	0.33	0.35	−1.75
101	WeFold	14	2.94	2.38	8.16	5.63	4.78	0.32	0.34	−9.59
388	ProQ2	15	2.11	1.11	8.25	7.48	4.74	0.32	0.32	20.29
330s	BAKER-ROSETTASERVER	15	2.63	2.57	7.26	5.78	4.56	0.30	0.30	29.53

The columns labeled GDT-HA, GDC-all, RPF, and LDDT are the sum of Z scores across all models submitted by each predictor group. The Sum column is the average of the sum of Z scores for the four metrics assessed. The Avg-s and Avg-a scores are the Sum scores divided by the number of AUs for which a model was submitted (s) by each predictor group, and the total number of AUs used for assessment (a), respectively. These scores are identical for predictor groups who submitted models for all TBM_Hard hAUs.

other three scores. Third ranked Mufold (197) had a zero GDC-all count. It appears that TASSER and RaptorX-ZY may have some part of the core packing matched with some experimental structures, yielding higher GDC-all scores even though the main chains of these models may not be more accurate than other models from other groups. However, global alignment scores like GDC-all are challenged when the overall backbone structure is inaccurate, suggesting that GDT-HA and GDC-all scores may be unreliable for assessing predictions of difficult TBM_hard AUs. Despite some of the valuable features of the superimposition-independent metrics like RPF and LDDT for difficult targets like TBM_hard AUs, human judgment like that used in the CASP10 FM assessment[14] is still a better approach than using automated scoring metrics.

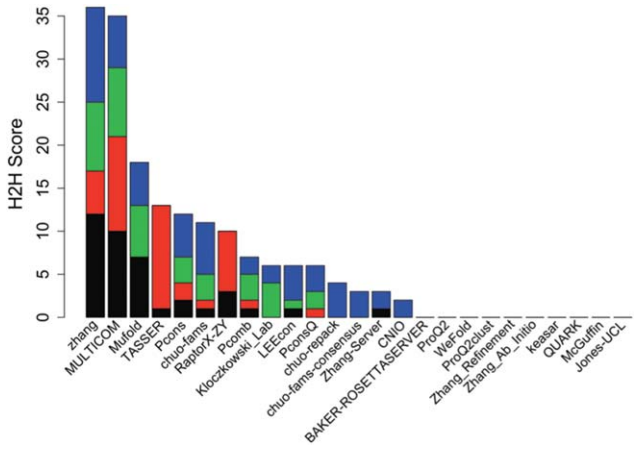

Figure 7
Ranking of top 25 human/server predictor groups on TBM_hard targets. Head-to-head (H2H) pairwise Student's *t*-test analysis on raw scores between 25 top-ranking server/human predictor groups for 15 TBM_hard AUs. Black, GDT-HA; red, GDC-all; green, RPF-9; blue, LDDT-15. Top-ranking 25 groups were identified based on the Avg-a *Z* score.

Impact of accurate model selection

As discussed above in the H2H assessment of servers, a few poor models can make a significant impact on the relative ranking, particularly in distinguishing the top-performing group from other groups that generally did very well in CASP10. In many cases, predictor groups included in their five submitted models more accurate models than the one designated as Model 1. This is a well-recognized aspect of previous CASP experiments. Interestingly, our analysis suggests that at least in some cases, the more accurate model could have been selected using alternative fold-accuracy discriminators. Figure 8 illustrates two

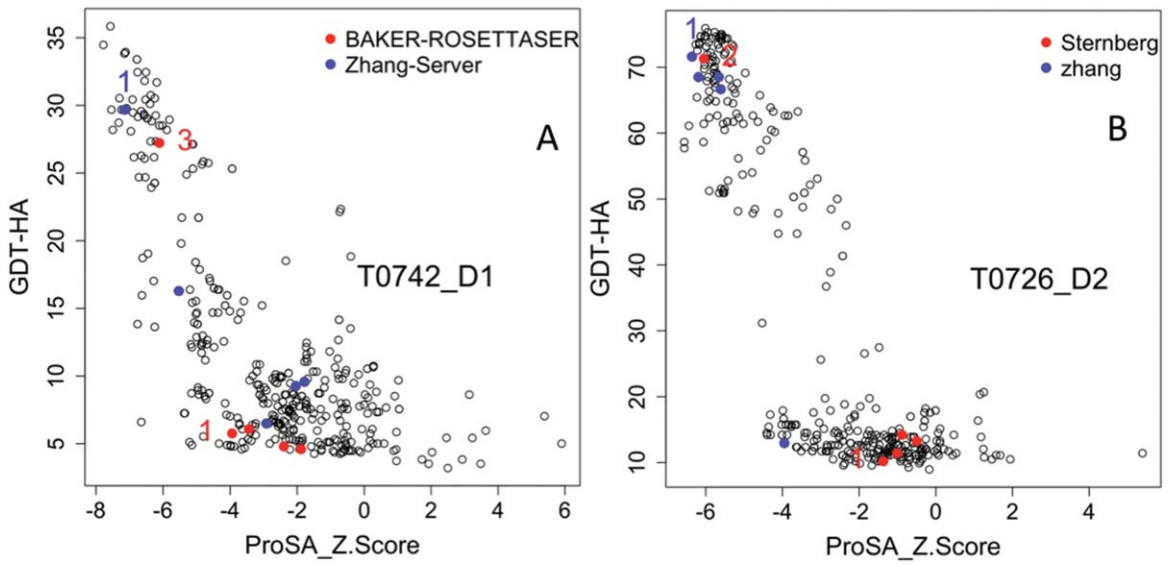

Figure 8

The challenge of accurate model selection. Predictor groups often submitted a much less accurate model as Model 1 compared to other of their own submitted alternative models. **A**: For AU target T0742_D1, Model 1 submitted by group 035s (Zhang-Server) has much higher GDT-HA score (∼0.3; *y* axis) than Model 1 submitted by server group 330s (BAKER-ROSETTASERVER; GDT-HA ∼ 0.05). However, alternative Model 3 of group 330s (GDT-HA ∼ 0.28) is a more accurate model. **B**: For AU target T0726_D2, Model 1 submitted by the human-curated group 237 (Zhang) had much higher GDT-HA score (∼0.72) than Model 1 submitted by group 458 (Sternberg; GDT-HA ∼ 0.12). However, alternative Model 2 of group 458 (GDT-HA ∼ 0.71) is a much more accurate model. In these particular cases, the Prosa *Z* score (*x* axis) could have provided a good criterion in model selection, as it would have indicated a much more accurate "best model" for groups 330s and 458.

examples, one for server BAKER-ROSETTASERVER (330s) and one for human-curated group Sternberg (458) where an alternative model is more accurate (higher GDT-HA raw score) than the designated Model 1. In these two cases, the more accurate model also has a significantly better ProsaII[23] fold score.

We also did a comparison of GDT-HA scores between models picked from among submitted models by ProsaII score only, and the Model 1 provided by predictors for all combinations of Groups and AUs for which ProsaII scores are available from the Prediction Center. If all the groups used ProsaII to pick their best model from the five submitted, 91 groups would improve their average GDT-HA score. Our point is not that ProsaII is the single best indicator of model quality. However, as is generally appreciated, many CASP10 participants would benefit by more successful model selection.

Ranking based on ideal model selection

Selection of the most-accurate model among submitted alternate models is an important area for development to provide more accurate TBM predictions. These trends were validated by simulating an "ideal model selection" CASP10 competition, choosing for each group and AU the single model with highest GDT-HA as Model 1, and repeating the *Z* score and H2H raw score rankings as outlined in the previous sections. Rankings based on H2H Student's *t*-test raw score comparisons are shown in Figure 9(A) for server predictors and Figure 9(B) for human-curated predictors. While "ideal model selection" has modest impact on server rankings, it has a significant impact on human-curated predictor rankings (cf. the rankings of predictors in Figures 4 and 9). Server predictors for which ranking was significantly improved in this simulated ideal model selection CASP10 competition included BAKER-ROSETTASERVER (330s), PconsM (103s), MULTICOM-CLUSTER (081s), and MULTICOM-REFINE (125s). Human-curated predictors for which ranking was significantly improved by simulated ideal model selection include CNIO (475) [which using ideal model selection ranked second behind Zhang (237)], Baker (477), chuo-repack (294), chuo-fams (365), Zhang-refinement (490), and keasar (315). These results demonstrate the key role of model selection in accurate TBM protein structure prediction.

CONCLUSIONS

The CASP10 TBM assessment identified the Zhang-Server (035s), QUARK (114s), PMS (108s), Leecon (027), and Zhang (237) groups as providing the most accurate models for the AU targets. In the course of our assessment, we observed many examples where raw scores were more discriminating for distinguishing structural accuracy than *Z* scores (also see Supporting Information Figure S6), or where *Z* scores suggest good

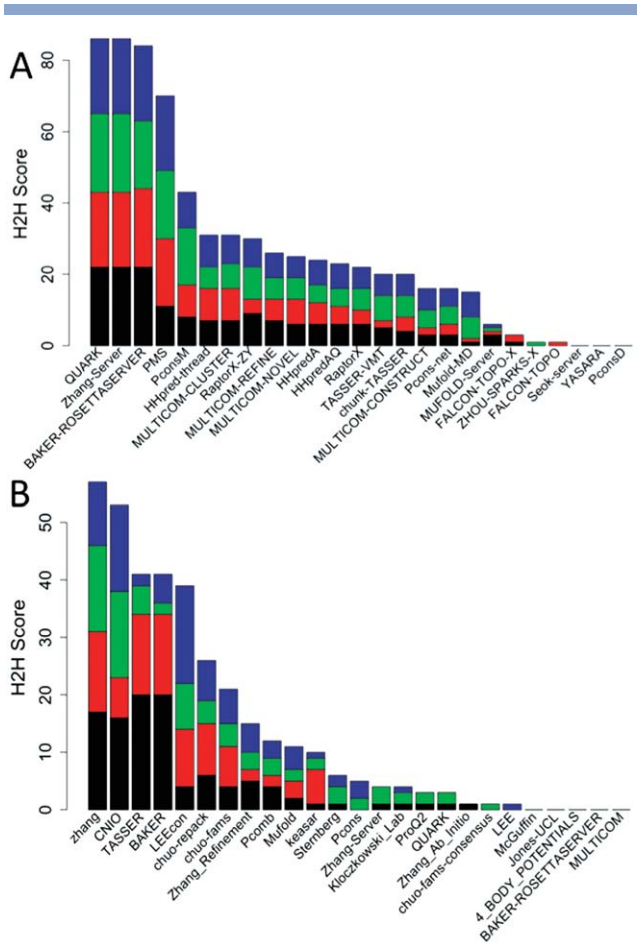

Figure 9

Ideal model selection. For each predictor group, the submitted model with highest GDT-HA score was selected and the paired Student's *t*-test analysis was done as in Figure 4. **A**: Head-to-head paired Student's *t*-test analysis on raw scores between 25 top-ranking server predictor groups for 112 hsAUs. **B**: Head-to-head paired Student's *t*-test analysis on raw scores between 25 top-ranking server/human predictor groups for 57 hAUs. Black, GDT-HA; red, GDC-all; green, RPF-9; blue, LDDT-15. Top-ranking 25 groups were identified based on the Avg-a *Z* score.

prediction performance while the raw scores together with visual inspection reveal inaccurate models (e.g., Fig. 3). In our experience, rankings based on raw scores, using head-to-head (H2H) comparisons of performance on common AUs, with a statistical assessment of whether one group outperforms each of the other groups, provides a more sensitive automated numerical assessment than *Z* scores alone. The H2H ranking developed by Mariani *et al.*[1] incorporates Student's *t*-test, providing a statistically sound basis for ranking the relative performance of groups on the same set of targets.

The CASP10 TBM assessment, evolving from the previous CASP assessment protocols, put heavy emphasis on side-chain atom positions, which are assessed by three (GDC-all, LDDT-15, and RPF-9) of the four structure accuracy metrics that contribute to our composite score. The philosophy was to assess TBM models with the same kinds of expectations for physically reasonable structures as is expected for experimental NMR or X-ray crystal structures. We also combined both global alignment scores (GDT-HA and GDC-all) which utilized superimpositions, and distance-matrix–based methods (LDDT-15 and RPF-9) which do not involve superimposition methods and are more sensitive to local structure accuracy and core side-chain packing features. The distance-matrix–based methods seem to be less sensitive to the loosely packed loop regions and also to helix tilt angles. Normalizing the RPF score against a random coil model impacts the ranking for the difficult targets. These multiple scores incorporate information on all heavy atom (C, N, O, and S) positions, and balance biases inherent in superimposition-based structural comparisons (e.g., GDT scores) which are problematic when comparing structures lacking high overall structural similarity.

ACKNOWLEDGMENTS

The authors thank all the scientists who contributed experimental structures to the CASP10 project, without which this project would not have been possible. Special thanks to A. Kryshtafovych for extensive efforts to support our study by providing guidance based on historical CASP assessment projects, for implementing new programs on the CASP Prediction Center cpu cluster, and for running the extensive numerical analyses that were the basis for our assessment. They also thank B.K. Lee, C.-H. Tai, H. Bai, J. Block, K. Fidelis, J. Moult, T. Schwede, T.J. Taylor, and A. Tramontano for extensive scientific discussions and helpful criticisms.

REFERENCES

1. Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T. Assessment of template based protein structure predictions in CASP9. Proteins 2011;79 (Suppl 10):37–58.
2. 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.
3. Moult J. Comparative modeling in structural genomics. Structure 2008;16:14–16.
4. Liu J, Montelione GT, Rost B. Novel leverage of structural genomics. Nat Biotechnol 2007;25:849–851.
5. Nair R, Liu J, Soong TT, Acton TB, Everett JK, Kouranov A, Fiser A, Godzik A, Jaroszewski L, Orengo C, Montelione GT, Rost B. Structural genomics is the largest contributor of novel structural leverage. J Struct Funct Genom 2009;10:181–191.
6. Montelione GT. The protein structure initiative: achievements and visions for the future. F1000 Biol Reports 2012;4:7.
7. Zhang Y, Thiele I, Weekes D, Li Z, Jaroszewski L, Ginalski K, Deacon AM, Wooley J, Lesley SA, Wilson IA, Palsson B, Osterman A, Godzik A. Three-dimensional structural view of the central

metabolic network of *Thermotoga maritima*. Science 2009;325:1544–1549.
8. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, Maniatis T, Califano A, Honig B. Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature 2012;490:556–560.
9. Huang YJ, Hang D, Lu LJ, Tong L, Gerstein MB, Montelione GT. Targeting the human cancer pathway protein interaction network by structural genomics. Mol Cell Proteomics 2008;7:2048–2060.
10. Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read RJ, Baker D. High-resolution structure prediction and the crystallographic phase problem. Nature 2007;450:259–264.
11. Raimondo D, Giorgetti A, Bosi S, Tramontano A. Automatic procedure for using models of proteins in molecular replacement. Proteins 2007;66:689–696.
12. Terwilliger TC, Dimaio F, Read RJ, Baker D, Bunkoczi G, Adams PD, Grosse-Kunstleve RW, Afonine PV, Echols N. Phenix.mr_rosetta: molecular replacement and model rebuilding with Phenix and Rosetta. J Struct Funct Genom 2012;13:81–90.
13. Keedy DA, Williams CJ, Headd JJ, Arendall WB III, 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.
14. Tai C-H, Bai H, Taylor TJ, Lee BK. Assessment of template free modeling in CASP10 and ROLL. Proteins 2014;82(Suppl 2):57–83.
15. 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.
16. Kryshtafovych A, Fidelis K, Tramontano A. Evaluation of model quality predictions in CASP9. Proteins 2011;79 (Suppl 10):91–106.
17. Taylor TJ, Tai C-H, Huang YJ, Block J, Bai H, Kryshtafovych A, Montelione GT, Lee BK. Definition and classifcation of evaluation units for CASP10. Proteins 2014;82(Suppl 2):14–25.
18. Snyder DA, Grullon J, Huang YJ, Tejero R, Montelione GT. The expanded FindCore method for identification of a core atom set for assessment of protein structure prediction. Proteins 2014;82(Suppl 2):219–230.
19. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP Prediction Center infrastructure and evaluation measures in CASP10 and CASP ROLL. Proteins 2014;82(Suppl 2):7–13.
20. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
21. Chen VB, Arendall WB III, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr Sect D, Biol Crystallogr 2010;66:12–21.
22. Lovell SC, Davis IW, Arendall WB III, de Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins 2003;50:437–450.
23. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins 1993;17:355–362.
24. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 2013;29:2722–2728.
25. Huang YJ, Powers R, Montelione GT. Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. J Am Chem Soc 2005;127:1665–1674.
26. Rosato A, Aramini JM, Arrowsmith C, Bagaria A, Baker D, Cavalli A, Doreleijers JF, Eletsky A, Giachetti A, Guerry P, Gutmanas A, Güntert P, He Y, Herrmann T, Huang YJ, Jaravine V, Jonker HR, Kennedy MA, Lange OF, Liu G, Malliavin TE, Mani R, Mao B, Montelione GT, Nilges M, Rossi P, van der Schot G, Schwalbe H, Szyperski TA, Vendruscolo M, Vernon R, Vranken WF, de Vries S, Vuister GW, Wu B, Yang Y, Bonvin AM. Blind testing of routine, fully automated determination of protein structures from NMR data. Structure 2012;20:227–236.
27. Huang YJ, Rosato A, Singh G, Montelione GT. RPF: a quality assessment tool for protein NMR structures. Nucleic Acids Res 2012;40:W542–W546.
28. Flory PJ. Statistical mechanics of chain molecules. New York: Interscience Publishers; 1969.
29. Read RJ, Chavali G. Assessment of CASP7 predictions in the high accuracy template-based modeling category. Proteins 2007;69 (Suppl 8):27–37.
30. 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.
31. 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.
32. Tress ML, Ezkurdia I, Richardson JS. Target domain definition and classification in CASP8. Proteins 2009;77 (Suppl 9):10–17.
33. 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.
34. Kryshtafovych A, Krysko O, Daniluk P, Dmytriv Z, Fidelis K. Protein structure prediction center in CASP8. Proteins 2009;77 (Suppl 9):5–9.
35. Cozzetto D, Kryshtafovych A, Tramontano A. Evaluation of CASP8 model quality predictions. Proteins 2009;77 (Suppl 9):157–166.
36. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. Proteins 2003;53 (Suppl 6):352–368.