PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Evaluation of CASP8 model quality predictions

Domenico Cozzetto,[1*†] Andriy Kryshtafovych,[2†] and Anna Tramontano[1,3]

[1] Department of Biochemical Sciences, Sapienza – University of Rome, Rome 00185, Italy
[2] Genome Center, University of California, Davis, CA 95616
[3] Istituto Pasteur – Fondazione Cenci Bolognetti, Sapienza – University of Rome, Rome 00185, Italy

ABSTRACT

The model quality assessment problem consists in the *a priori* estimation of the overall and per-residue accuracy of protein structure predictions. Over the past years, a number of methods have been developed to address this issue and CASP established a prediction category to evaluate their performance in 2006. In 2008 the experiment was repeated and its results are reported here. Participants were invited to infer the correctness of the protein models submitted by the registered automatic servers. Estimates could apply to both whole models and individual amino acids. Groups involved in the tertiary structure prediction categories were also asked to assign local error estimates to each predicted residue in their own models and their results are also discussed here. The correlation between the predicted and observed correctness measures was the basis of the assessment of the results. We observe that consensus-based methods still perform significantly better than those accepting single models, similarly to what was concluded in the previous edition of the experiment.

Proteins 2009; 77(Suppl 9):157–166.

Key words: CASP; model quality assessment; structure prediction.

INTRODUCTION

Molecular modeling aims at providing a structural framework to elucidate protein function. This process is usually hypothesis-driven: researchers generate 3D models to interpret available experimental data and to design further assays to validate them. Over the past years, the improvement in the scope and reliability of protein structure prediction has been paralleled by the growing number and variety of its valuable applications in synergy with wet lab experiments. Computational models—typically derived from template-based techniques—are now commonly used to identify key functional amino acids and can contribute to more advanced analyses, including drug discovery, dissection of protein interactions, and protein engineering, just to mention a few applications.[1,2]

A wealth of easy-to-access prediction tools and model repositories[3–5] abound now on the Internet. However, making the most effective use of 3D models inevitably requires trustworthy *a priori* estimates of their divergence from the native conformation. Nowadays this information is rarely supplied to the end users, who are left to decide by themselves whether a model is suitable for their specific problem.

Effective tools for this task would be instrumental for the development and improvement of fold recognition methods, fragment assembly approaches, and meta-predictive systems. In response to these needs, the bioinformatics community has focused on the Model Quality Assessment (MQA) problem, i.e., on the possibility to predict the global and local accuracy of 3D models when experimental structural data are not yet available. More than twenty articles have been published on the subject in the last 3

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

This article is dedicated to the memory of our friend and colleague Angel Ortiz.

The authors state no conflict of interest.

Grant sponsor: KAUST; Grant number: KUK-I1-012-43; Grant sponsor: MIUR (FIRB Rete Italiana di Proteomica), EMBO; Grant sponsor: The US National Library of Medicine; Grant number: LM-7085

†These two authors equally contributed to this work.

*Correspondence to: Domenico Cozzetto, Department of Biochemical Sciences "Rossi Fanelli", Sapienza – University of Rome, Rome 00185, Italy. E-mail: domenico.cozzetto@uniroma1.it.

Received 14 April 2009; Revised 19 June 2009; Accepted 25 June 2009
Published online 14 July 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22534

years.[6] Thus far, physics-based energies, statistical potentials, and machine-learning techniques have been applied to this challenge with various degrees of success. When several independent models for the same target protein are accessible, a consensus approach can be exploited by scoring each prediction according to its similarity to the whole collection.

As in other research areas, the objective evaluation of such predictions is an essential step to identify effective strategies and assist further progress. The CASP7 organizers recognized the importance of this issue and launched MQA as a new category of critical judgement in 2006.[7] The Prediction Center website publicly released the server models submitted for assessment in the tertiary structure prediction categories. Predictors were invited to download such models, generate quality estimates and submit them to CASP before the corresponding experimental structures were available—according to the deadlines set by the organizers. In CASP8, predictions were also accepted in a server regime: per-target tarballs containing the 3D server models were automatically submitted to the registered servers that had three calendar days to respond with their estimates. The prediction format for the overall reliability of models (QM1) required scores as real numbers ranging from 0.0 to 1.0—where a higher score corresponds to an estimated better model. A score of 1.0 should correspond to a perfect model. For the per-residue accuracy analysis, scores should report distances in Angstroms between the corresponding residue Cα atoms after a sequence-dependent superposition of the 3D model to the native structure.

CASP8 demonstrated growing interest in the development of methods for model accuracy prediction, as the number of participants almost doubled from the previous experiment. Forty-five groups (including 30 servers) submitted predictions for QM1 and 17 of them also specified local confidence values (QM2). Here, we assess the performance of these groups, as well as that of the 83 groups participating in the tertiary structure prediction category and estimating the local correctness of their own models (QM3).

The assessment described here is based on the correlation between the predicted and experimental accuracy values for each group, and on the statistical significance of the observed differences. It should be mentioned that the CASP8 targets were divided into two groups[8]: human/server targets (targets meant for structure prediction by all participating groups) and server only targets. In the MQA category, groups were invited to submit predictions for both types of targets.

MATERIALS AND METHODS

The Protein Structure Prediction Center website provides open access to:

1. the server models—35,882 altogether for 121 CASP8 targets[9]—released for quality prediction—at http://www.predictioncenter.org/download_area/CASP8/server_predictions/
2. the overall and residue-based reliability estimates submitted by participating groups—at http://predictioncenter.org/download_area/CASP8/predictions/QA.tar.gz and http://predictioncenter.org/casp8/qa_analysis.cgi
3. the global and local accuracy scores in terms of GDT-TS values and distances between the corresponding residues in the target and the model resulting from optimal LGA[10] superposition—at http://www.predictioncenter.org/casp8/results.cgi
4. the results of the correlation analysis between the predicted and observed deviation of the models from the experimental structures—at http://predictioncenter.org/casp8/qa_analysis.cgi
5. the 3D models submitted for targets assigned to the TBM category—at http://predictioncenter.org/download_area/CASP8/predictions/
6. the list of group IDs used for blind assessment and the corresponding group names—at http://predictioncenter.org/casp8/docs.cgi?view=groupsbynumber

The authors of Ref. 11 kindly supplied their estimates for the modeling difficulty of each target.

Detailed statistical analyses of the results were performed using in-house R[12] scripts.

The application of Fisher's transformation preceded the statistical comparison between two correlation coefficients coherently with standard statistical practice.[13] The equation

$$z' = (\ln(1+r) - \ln(1-r))^{*}0.5$$

defines the relationship between Pearson's r and a normally distributed variable z' with variance $s^2 = 1/(n-3)$, where n is the number of observations. Two correlation coefficients r_1 and r_2 can be converted into the corresponding z'_1 and z'_2, whose difference is normally distributed with variance $s^2 = 1/(n_1-3) + 1/(n_2-3)$—where n_1 and n_2 represent the number of models evaluated by the two predictors. The p-value associated with $|z'_1 - z'_2|$ in the standard normal probability table is an estimate of the likelihood that the difference between r_1 and r_2 is statistically significant.

BLAST/LGA[14] is a naïve predictor that assigns a confidence index to a model on the basis of its structural divergence from the most closely related known protein structure detectable by standard sequence analysis. Specifically, it first searches the nr protein database[15]—frozen at the release date of the corresponding target—for the sequence of the protein of known structure that is most similar to the target by running at most five PSI-

BLAST[16] iterations with default parameters. Then it superimposes the selected structure onto the input protein model by running LGA with default parameters in sequence independent mode. Finally, the resulting LGA_S score is divided by 100 to obtain a number between 0.0 and 1.0.

RESULTS AND DISCUSSION

The correlation between the predicted and observed correctness values forms the basis of the scoring functions adopted to rank the groups. Ideally, prediction methods should give estimates that are linearly correlated to the experimentally observed accuracies of the models. Under such assumptions, Pearson's r coefficient is a sensible choice to evaluate group performance. However, Pearson's r also assumes normally distributed data—which is not always the case—hence, distribution-free measures such as Spearman's ρ or Kendall's τ could be preferable. We analyzed the data using both parametric and nonparametric inferential statistical methods and verified that choice of the association measure has only marginal effects on the evaluation results (Table SI in Supporting Information). In the following, we use Pearson's r for data analysis.

QM1: global accuracy of models

The distributions of Pearson's r were first calculated for each target separately to test the ability of methods to predict quality on a relative—i.e., protein-dependent scale. Target-based r distributions of 108/121 turned out to be normal after visual inspection and quantitative Shapiro-Wilk tests[17] at the 0.01 significance level.

The results obtained by each group for each target were converted into Z-scores. Negative values were set to zero, and the performance of each group was measured by the average of such modified Z-scores. The choice of neglecting negative Z-scores is meant not to penalize groups that, by attempting novel and riskier methods, might obtain negative scores in some cases. We verified that the overall conclusions were not affected by this choice.

Figure 1 shows the scores for all 45 predictors on the whole set of targets, while Supporting Information Table SII reports the average correlation coefficients for each group assessed. Different from CASP7, there is no clear gap between the scores of the better performing methods and the others, the distribution being rather smooth. The paired Student's t-test on the common set of predicted models (Table I) was used to assess the statistical significance of the observed differences between groups. The top-ranked four predictors (239 - Pcons_Pcons, 31 - ModFOLDclust, 56 - SAM-T08-MQAC, and 27 - QMEANclust) appeared to perform better than the remaining ones and to be indistinguishable from each other.

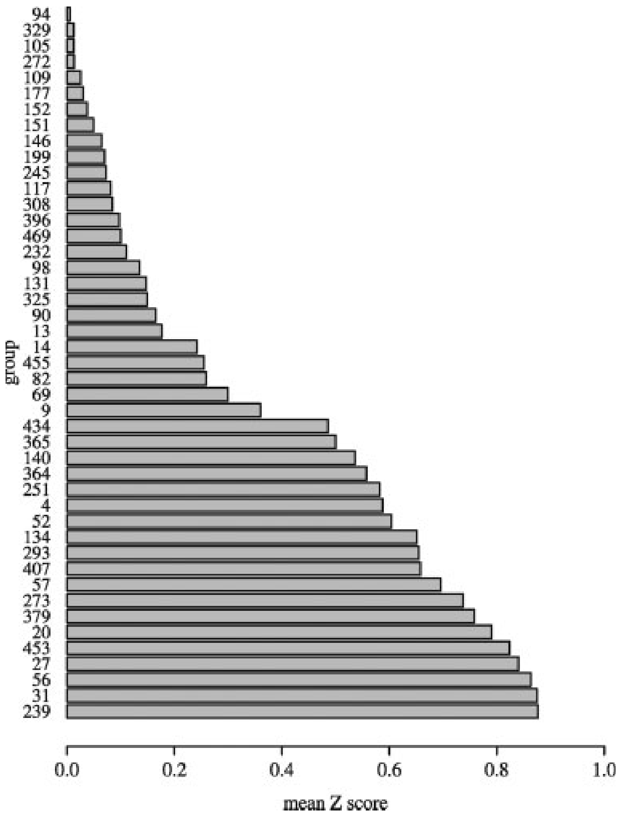

Figure 1

Scores for individual groups in the QM1 target-based prediction category.

The next step was to analyze the performance of the groups for the TBM and FM target categories separately. The CASP target classification is obviously domain-based while quality predictions are assigned to whole protein chains, therefore, we took into consideration 114 targets that either are single domain or whose domains belong to the same prediction category. We included in the TBM class the targets that were in both the TBM and FM category.[18] The breakdown of the predictor performance by structure prediction category reveals that the ranking in Figure 1 basically mirrors the one for the TBM targets (data not shown). This is easily explainable since there are only four FM targets in the resulting dataset; for this reason no conclusion can be derived about putative differences between the performance of methods for the FM category.

The target-by-target analysis of the correlation described above does not reflect the predictor ability to assign an absolute estimate to a model that would permit to compare the expected quality of models for different proteins. For this purpose, we repeated the analysis by pooling all models together. In the context of the global

Table I
Statistical Comparisons Among the Top 12 Groups Whose Global Quality Estimates Were Assessed on a Per Target Basis

	239	31	56	27	453	20	379	273	57	407	293	134
239		121 35,808	120 34,815	121 35,177	120 34,861	121 35,116	120 35,203	118 34,045	120 22,040	119 35,231	77 23,082	117 34,065
31	0.001		120 34,839	121 35,183	120 34,861	121 35,116	120 35,228	118 34,052	120 22,041	119 35,255	77 23,101	117 34,072
56	0.004	−0.004		120 34,198	120 34,165	120 34,148	120 34,541	117 33,207	120 21,755	119 34,559	76 22,302	116 33,230
27	0.014	0.014	0.016		120 34,311	121 34,565	120 34,592	118 33,546	120 21,885	119 34,608	77 22,606	117 33,569
453	−0.024	−0.024	−0.026	−0.009		120 34,844	120 34,551	117 33,267	120 21,816	119 34,594	76 22,437	116 33,286
20	0.035	0.034	0.036	0.022	0.01		120 34,534	118 33,518	120 21,816	119 34,577	77 22,706	117 33,537
379	−0.031	−0.032	−0.028	−0.015	0.005	0.005		117 33,458	120 21,835	119 34,942	76 22,809	116 33,478
273	−0.034	−0.035	−0.035	−0.023	0.015	−0.002	0.006		117 21,172	116 33,491	74 21,860	117 33,794
57	0.048	−0.05	−0.053	−0.037	0.023	−0.008	0.034	0.039		119 21,862	76 13,923	116 21,195
407	−0.068	−0.069	−0.065	−0.047	0.033	−0.023	−0.037	−0.023	−0.02		76 22,819	115 33,511
293	−0.063	−0.063	−0.058	−0.045	0.028	−0.027	0.045	−0.037	−0.018	0.012		74 21,862
134	0.056	−0.056	−0.058	−0.045	0.035	−0.024	0.027	0.022	0.013	−0.004	−0.017	

Results of the paired *t*-tests. The upper right part of the table contains the numbers of common targets and models, respectively. Estimated differences in the means of Pearson's correlation coefficients are in the lower left half. Gray cells highlight pairs of statistically indistinguishable groups at the 10^{-2} significance level.

correlation analysis, we followed the well-established procedure to assess the statistical significance of the difference between two correlation coefficients, by making use of their Fisher's z' transformation. Figure 2 reports the results of this study and indicates that Group 31 (ModFOLDclust) outperforms all the others, including the next best scoring ones 56 (SAM-T08-MQAC), 27 (QMEANclust), 453 (MULTICOM), and 239 (Pcons_Pcons). This difference is statistically significant (Table II).

For each group, we also assessed the difference in quality between the top-ranked model(s) for each target and the actual best one(s). In particular, for each single-domain target we computed the average GDT-TS difference (ΔGDT)[14] between the rank-one model(s) of a group and the most accurate one(s). Figure 3 reveals that, in general, the predicted best model might be rather far from the actual most reliable one. The mean loss in accuracy varies from 3.51 to 43.55 GDT-TS units for all predictors across all targets. Groups 239, 31, 56, 27, and 453 generally attain ΔGDT values that are lower than the average, yet none of them can consistently select the best model for all targets.

QM2: residue-level accuracy of models

Seventeen groups submitted confidence estimates at the residue level, and we evaluated the correlation between such values and the distances in Angstroms between the predicted and observed positions of each Cα after optimal sequence-based superposition of targets and models. Predictions for 77 models—approximately 0.21%

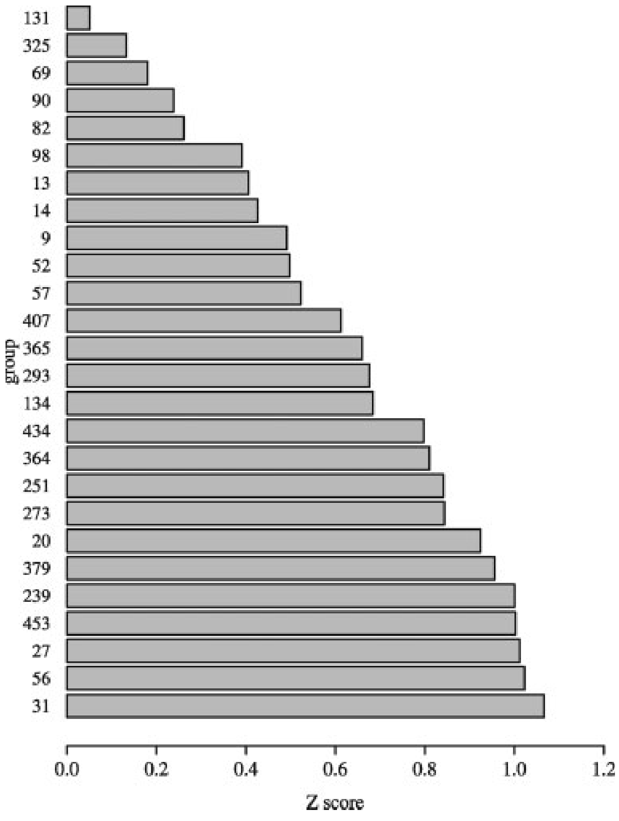

Figure 2
Scores for individual groups in the QM1 prediction category when the predictions for all targets are pooled together.

Table II

Statistical Comparisons Among the Top 12 Groups in the Global Assessment of their QM1 Predictions

	31	56	27	453	239	379	20	273	251	364	434	134
31	35,833											
56	−6.637	34,844										
27	8.186	1.522	35,183									
453	−9.51	−2.853	−1.338	34,861								
239	−9.837	−3.133	−1.608	0.261	35,808							
379	−15.831	−9.096	−7.589	6.231	−6.011	35,541						
20	19.782	13.04	11.547	10.184	9.991	3.991	35,116					
273	−28.848	−22.082	−20.619	19.242	−19.11	13.127	−9.132	34,347				
251	−29.02	−22.263	−20.802	19.427	−19.296	13.322	−9.332	0.212	34,151			
364	−30.174	−23.856	−22.49	21.203	−21.083	15.494	−11.758	−3.211	−3.009	26,662		
434	−33.842	−27.001	−25.542	24.146	−24.046	−18.02	−13.986	−4.776	−4.555	−1.234	35,138	
134	−44.211	−37.354	−35.93	34.521	34.49	28.491	−24.458	15.26	15.027	11.064	10.58	34,082

Results of the *Z* tests. Pearson's coefficients for all participants were computed and the distributions of their corresponding Fisher's z' compared. Diagonal entries contain the number of models for which the corresponding group submitted QM1 predictions. Values of the *z* statistics are shown in the lower left cells. Gray cells highlight pairs of statistically indistinguishable groups at the 10^{-2} significance level.

of the whole decoy set size—were submitted by less than 7 groups, so they were not considered further.

For each server model we computed the Pearson's *r* coefficients and the corresponding Z-scores. The final score of each predictor was determined by the average Z-score over the set of predicted models—after replacing negative values with 0s (Fig. 4). The statistical significance of the results was assessed by paired Student's *t*-tests on the common residues of the common models (Table III).

It is rather apparent that Group 27 (QMEANclust) outperformed all other predictors in this task.

As in QM1, we also analyzed the results separately for different target categories (data not shown) and the overall ranking, including the good performance of Group 27, did not change.

Benchmark calculations for QM1

Previous CASP editions proved that template-based techniques cannot improve consistently over the best available single template structure.[19–22] In other words, the overall reliability of comparative models is likely to decrease as they depart from the starting template structure. The BLAST/LGA naïve predictor exploits this observation and infers the global accuracy of a single TBM model as a function of its distance from the best template found by sequence searching

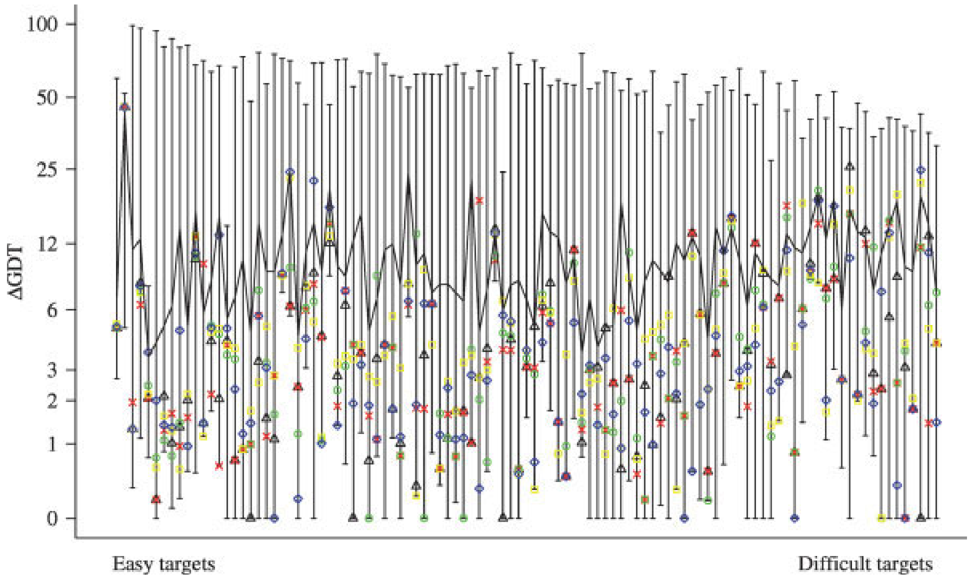

Figure 3

ΔGDT values as a function of the target modeling difficulty. For each target, vertical bars represent the range of ΔGDT values on a logarithmic scale, while the black segments connect the means. Results of groups 239 (black triangles), 31 (red crosses), 56 (green circles), 27 (yellow squares), and 453 (blue diamonds) are shown.

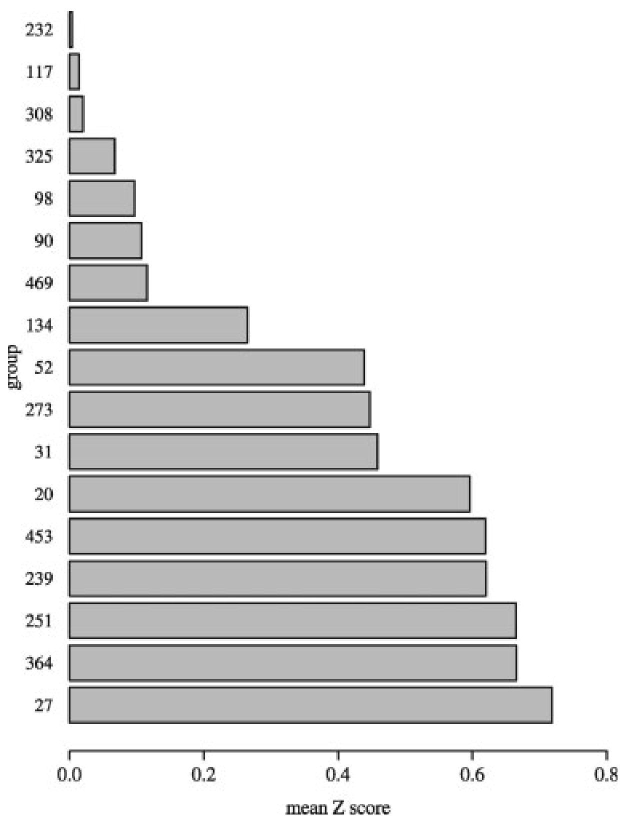

Figure 4
Scores for individual groups in the QM2 category.

approaches. For 56 single-domain targets, PSI-BLAST detected at least one similar sequence with known structure and BLAST/LGA produced 14,964 QM1 estimated values. To compare these naïve predictions with the official groups' ones in an unbiased manner, BLAST/LGA correlation values were excluded from the computation of the Z-scores. The Z-scores for this method were computed from the average and standard deviation values of the Pearson's *r* distributions for the 45 official predictors.

BLAST/LGA performs worse than the highest scoring methods, but is statistically indistinguishable from most of the remaining ones, as Figure 5 and Supporting Information Table SIII document.

The most effective methods in CASP7 were consensus-based approaches—such as Pcons[23]—and the assessment team highlighted the need to improve tools able to assign reliability scores to single models.[14] Once again, all top performing methods with the possible exception of Group 293 (LEE SERVER) rely on the identification of a consensus among the models. Yet, further investigations prove that such methods tend to achieve results similar to single model approaches as the 3D modeling difficulty of the corresponding target sequence increases (Supporting Information Figure S1).

QM3: self assessment of residue-level accuracy

In CASP, modeling groups have long had the possibility of labeling predicted residues with error estimates

Table III
Statistical Comparisons Among the Top 12 Groups Involved in the QM2 Test

	27	364	251	239	453	20	31	273	52	134	469	90
27		25,648 4,070,408	33,949 6,198,283	35,106 6,565,569	34,295 6,420,411	34,549 6,484,898	35,112 6,568,180	33,671 6,238,654	34,788 6,519,558	33,412 6,220,321	34,050 6,295,164	34,811 6,516,234
364	−0.021		25,479 4,074,294	26,190 4,190,792	25,428 4,018,366	25,683 4,083,117	26,178 4,156,097	25,433 4,023,282	25,854 4,142,180	25,171 4,004,835	25,452 4,045,303	25,530 4,079,660
251	−0.022	0.006		34,075 6,216,388	33,273 6,071,469	33,527 6,135,956	34,081 6,217,022	32,932 5,995,207	33,758 6,170,390	32,673 5,976,874	33,037 5,954,250	33,780 6,167,053
239	−0.04	0.023	0.019		35,111 6,641,580	35,366 6,706,331	36,029 6,811,422	34,576 6,475,790	35,705 6,817,926	34,314 6,457,343	34,920 6,575,809	35,262 6,723,376
453	−0.035	−0.013	−0.013	0.001		35,094 6,641,955	35,111 6,643,434	33,828 6,323,811	34,784 6,592,852	33,565 6,305,132	34,016 6,358,130	34,396 6,510,402
20	0.053	0.04	0.031	0.014	0.015		35,366 6,708,185	34,079 6,388,004	35,037 6,657,573	33,816 6,369,325	34,263 6,420,763	34,650 6,574,889
31	−0.083	0.065	0.061	0.043	0.045	−0.029		34,460 6,455,603	35,707 6,764,161	34,198 6,437,156	34,895 6,521,254	35,246 6,669,349
273	−0.099	0.067	−0.078	−0.059	0.061	−0.045	−0.017		34,293 6,427,197	34,328 6,459,062	33,577 6,220,435	33,893 6,349,656
52	−0.117	0.104	0.097	0.077	0.075	−0.06	−0.033	0.017		3,403,123 6,408,750	34,584 6,523,958	34,929 6,674,189
134	−0.184	0.157	0.164	0.146	0.149	−0.134	−0.104	0.087	−0.069		33,320 6,202,333	33,634 6,331,323
469	−0.301	−0.285	−0.282	−0.263	−0.263	−0.247	−0.221	−0.203	−0.187	−0.116		34,196 6,445,339
90	−0.293	0.279	0.274	0.253	0.256	−0.24	−0.21	0.191	−0.176	0.105	−0.014	

Results of the paired *t*-test. Pearson's coefficients for residues of common models were computed and their distributions compared. Cells in the upper left part of the table show the numbers of common models and residues, respectively. Estimated differences in the means of Pearson's *r* are in the lower left half. Gray cells highlight pairs of statistically indistinguishable groups at the 10^{-2} significance level.

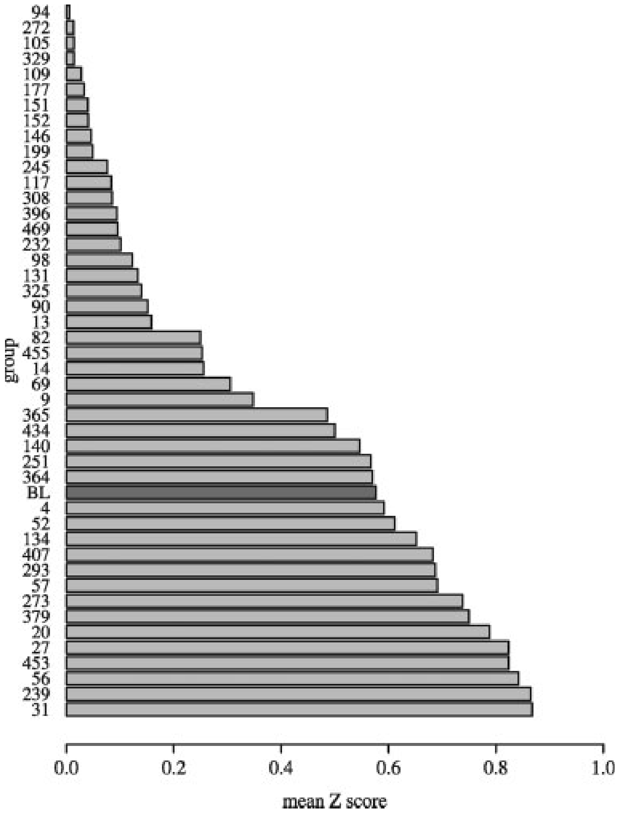

Figure 5

Comparison of the BLAST/LGA method (BL) with all prediction groups submitting global quality estimates for models corresponding to TBM targets.

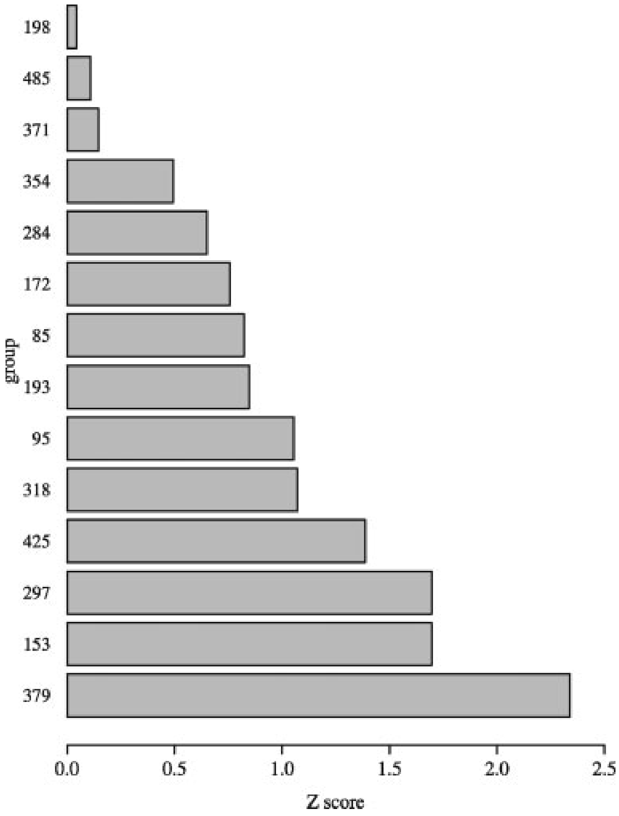

Figure 6

Scores for the prediction groups in the QM3 category where predictors provide residue-based error estimates for their own models. Only groups with a positive score are shown.

through the B-factor field in the TS format. In CASP7, the analysis of the QM3 results was included in the TBM assessment,[22] while in this edition it is described in this article.

Eighty-three groups submitted at least two different values in the B-factor column for more than 10 targets for a total of 1,333,541 residue-level error estimates. However, a high proportion of such Cα error estimates turned out to represent physically unrealistic distances. We applied a filter to the 3D predictions discarding those where less than 90% of the values in the B-factor column were in the range from 0.0 to 10.00. This step reduced

Table IV

Statistical Comparisons Among the Top 12 Groups Assessed in QM3

	379	153	297	425	318	95	193	85	172	284	354	371
379	14,973											
153	−3.408	542										
297	−3.408	0	542									
425	−20.234	1.493	1.493	21,944								
318	−22.59	2.914	2.914	−5.55	11,853							
95	−24.691	3.006	3.006	−6.395	−0.272	15,933						
193	29.583	3.935	3.935	10.855	3.734	3.764	19,569					
85	−28.808	4.03	4.03	−10.777	3.984	4.021	−0.433	16,363				
172	26.604	4.276	4.276	10.441	4.516	4.553	1.406	0.993	10,548			
284	−20.963	4.583	4.583	−8.743	4.584	4.559	−2.239	−1.929	−1.132	4435		
354	−7.64	3.606	3.606	−3.426	−2.157	−2.101	−1.312	−1.218	−0.967	−0.556	402	
371	−32.488	6.785	6.785	−18.074	11.965	12.391	−9.776	−9.176	−7.613	−4.955	1.245	7922

Results of the *Z* tests. Pearson's coefficients for all participants were computed, and the distributions of their corresponding Fisher's z' compared. Diagonal entries contain the number of residues in the filtered dataset for which the corresponding group submitted QM3 predictions. Values of the *z* statistics are shown in the lower left cells. Gray cells highlight pairs of statistically indistinguishable groups at the 10^{-2} significance level.

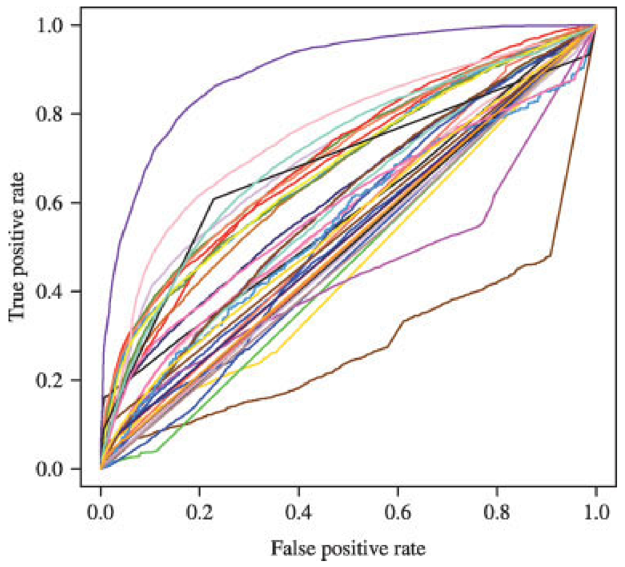

Figure 7

ROC curves comparing the 3D modeling groups' ability to identify well positioned residues in their own predictions. Group 379 (blue line) achieves the highest area under the curve.

the dataset size to 661,992 error estimates from 60 groups and corresponding to 152 TBM assessment units for which error estimates were submitted by at least ten independent predictors.

Groups were scored by the procedure adopted for the QM1 global correlation test and by receiver operator characteristic (ROC) curves. For the purpose of ROC analysis, we tested the predictors' ability to detect correctly modeled residues using the same strategy devised by the CASP7 TBM assessor.[22] We considered a modeled Cα atom to be correct, if it falls within 3.8 Å of the corresponding experimental position after optimal sequence-dependent LGA alignment. Next, for each group we scaled the estimated errors in [0.0, 1.0] and—by varying the discrimination threshold across such values—counted the number of true positives (TP—error estimates less than or equal to the current threshold and classified as correct), true negatives (TN—inferred distances greater than the present cut-off and classified as incorrect), false positives (FP—error estimates less than or equal to the actual threshold and classified as incorrect), and false negatives (FN—predicted misplacements that are greater than the cut-off but classified as correct). In Figure 9, we plot the false positive rate (FPR) versus the true positive rate (TPR), which are defined by the formulas: FPR = FP/(FP + TN) and TPR = TP/(TP + FN).

Group 379 (McGuffin) performed significantly better than the others in the correlation analysis (Fig. 6 and Table IV), and the ROC analysis (Fig. 7) confirms this finding.

Comparison between CASP8 and CASP7

The comparison of the CASP7 and CASP8 distributions of correlation coefficients is a basic step in estab-

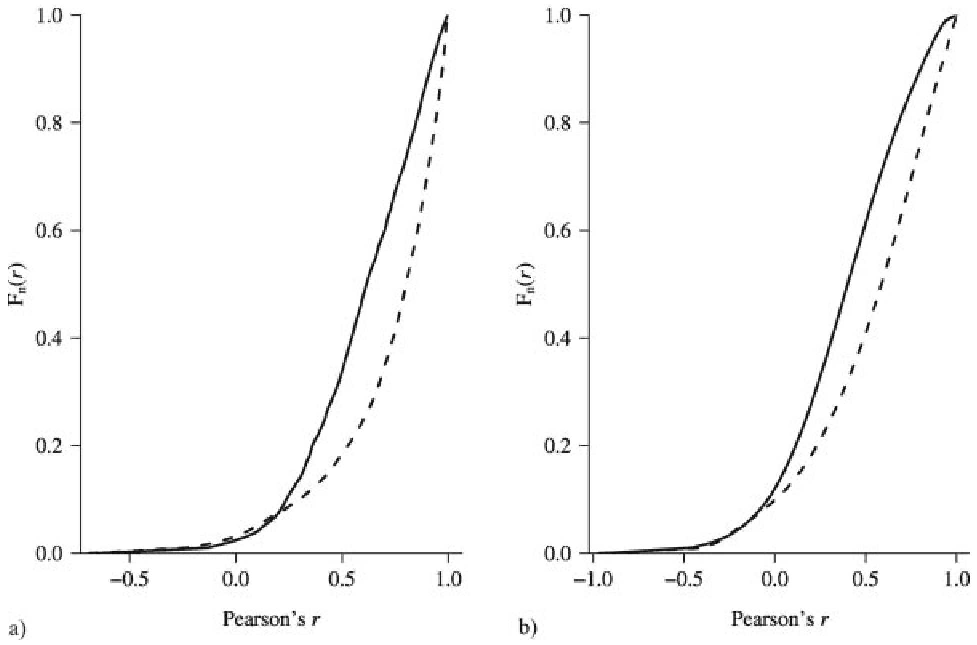

Figure 8

Empirical cumulative distribution functions of Pearson's *r* for the QM1 (**a**) and QM2 (**b**) experiments. Solid and dotted lines represent CASP7 and CASP8 data, respectively.

lishing whether the MQA field is improving. Figure 8 shows the results of this analysis in terms of empirical cumulative distribution functions that report the proportion of observed Pearson's *r* less than or equal to a given threshold. It is apparent that the fraction of cases where *r* is greater than 0.5 has largely increased for both the QM1 and QM2 experiments in CASP8 with respect to CASP7.

Such results definitely point to progress in the area, although the reasons behind it are not clear. Higher correlations in CASP8 might arise as a consequence of genuine method improvements, but they might also reflect a difference in the difficulty of the test sets used in the two experiments. Unfortunately, deriving a difficulty scale for an unbiased comparison of the MQA results in subsequent CASP rounds is intrinsically hard. As emerged from the first two assessments of this prediction category, the 3D structure of two targets might be roughly similarly difficult to predict[11] and yet the difficulty of ranking the accuracy of the collected server models might substantially differ. Indeed, Figure 4 and Supporting Information Figure S1 demonstrate that the accuracy of QM1 predictions only weakly correlates with the estimated difficulty of the target 3D prediction.

CONCLUSIONS

The results of the second round of MQA in CASP highlighted methods that are better in assigning error estimates at both the residue and global level, and it contributed to develop a robust and sound assessment procedure. Nevertheless, a few issues remain to be addressed.

The ability to rank models by consensus methods, that is, to sort a set of models for the same target according to their quality, has important applications in fold recognition and fragment-based methods for protein structure prediction, and is extremely useful for meta-predictor performance. However, it is of very limited usage to the end users of models, who need to be provided with an estimate value of a single model or of its regions that can be, in turn, used to identify the scope of application of the model itself. Similarly to what was the case in CASP7, we are forced to conclude that there is still no reliable tool for this purpose. Once again, therefore, we urge the community to concentrate their efforts on this important area.

Another open issue is the detection of progress in the field. When methods are compared on the same set of targets—as is the case in CASP—several statistical tools are available to comparatively evaluate the results, as shown in this article. How can we reckon the extent of the advancement made over time? The problem arises when the task is to compare different approaches on different test sets. At first sight, it might look reasonable to take into account the specific method used. As an example, for a consensus-based method, parameters such as the distribution and the cluster properties of the different models could be considered. However, this is still unsatisfactory when different methodologies need to be compared. This is one problem that the community should discuss before the next MQA experiment.

We would like to stress once more the relevance of the MQA experiments in all their flavors. It is one area of protein structure prediction where success benefits both developers and users and is particularly important to encourage structure prediction groups to provide quality estimates with their models. We look forward to seeing more and more protein structure modeling resources including accuracy estimates with their results.

ACKNOWLEDGMENTS

The authors are grateful to the template-based modeling CASP7 assessment team for valuable help and discussions.

REFERENCES

1. Tramontano A. The role of molecular modelling in biomedical research. FEBS Lett 2006;580:2928–2934.
2. Schwede T, Sali A, Honig B, Levitt M, Berman HM, Jones D, Brenner SE, Burley SK, Das R, Dokholyan NV, Dunbrack RL, Fidelis K, Fiser A, Godzik A, Huang YJ, Humblet C, Jacobson MP, Joachimiak A, Krystek SR, 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. Kiefer F, Arnold K, Kunzli M, Bordoli L, Schwede T. The SWISS-MODEL repository and associated resources. Nucleic Acids Res 2009;37:D387–D392.
4. Pieper U, Eswar N, Webb BM, Eramian D, Kelly L, Barkan DT, Carter H, Mankoo P, Karchin R, Marti-Renom MA, Davis FP, Sali A. MODBASE, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 2009;37:D347–D354.
5. Castrignano T, De Meo PD, Cozzetto D, Talamo IG, Tramontano A. The PMDB Protein Model Database. Nucleic Acids Res 2006;34:D306–D309.
6. Kryshtafovych A, Fidelis K. Protein structure prediction and model quality assessment. Drug Discov Today 2009;14:386–393.
7. Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction-Round VII. Proteins 2007;69(Suppl 8):3–9.
8. Kryshtafovych A, Krysko O, Daniluk P, Dmytriv Z, Fidelis K. Protein Structure Prediction Center in CASP8. Proteins 2009;77(Suppl 9):5–9.
9. Tress ML, Ezkurdia I, Richardson JS. Target domain definition and classification in CASP8. Proteins 2009;77(Suppl 9):10–17.
10. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
11. Kryshtafovych A, Fidelis K, Moult J. CASP8 results in context of previous experiments. Proteins 2009;77(Suppl 9):217–228.
12. The R Development Core Team. R: A language and Environment for statistical computing, Vienna: R Foundation for Statistical Computing; 2006.

13. Sheskin DJ. Handbook of Parametric and Nonparametric Statistical Procedures, Boca Raton: Chapman & Hall/CRC; 2007. p 1776.
14. Cozzetto D, Kryshtafovych A, Ceriani M, Tramontano A. Assessment of predictions in the model quality assessment category. Proteins 2007;69(Suppl 8):175–183.
15. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Yaschenko E, Ye J. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2009;37:D5–D15.
16. 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.
17. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika 1965;52(3–4):591–611.
18. Ezkurdia I, Richardson JS, Tress M. Domain definition and target classification for CASP8. Proteins 2009, this issue.
19. Tramontano A, Leplae R, Morea V. Analysis and assessment of comparative modeling predictions in CASP4. Proteins 2001;(Suppl 5)22–38.
20. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. Proteins 2003;53(Suppl 6):352–368.
21. Tress M, Ezkurdia I, Grana O, Lopez G, Valencia A. Assessment of predictions submitted for the CASP6 comparative modeling category. Proteins 2005;61(Suppl 7):27–45.
22. 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.
23. Wallner B, Elofsson A. Prediction of global and local model quality in CASP7 using Pcons and ProQ. Proteins 2007;69(Suppl 8):184–193.