Assessment of predictions in the model quality assessment category

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

[1] Department of Biochemical Sciences, University of Rome "La Sapienza", P. le A. Moro, 00185 Rome, Italy
[2] Protein Structure Prediction Center, Genome Center, University of California, Davis, California
[3] Istituto Pasteur–Fondazione Cenci Bolognetti, University of Rome "La Sapienza", P. le A. Moro, 00185 Rome, Italy

ABSTRACT

The article presents our evaluation of the predictions submitted to the model quality assessment (QA) category in CASP7. In this newly introduced category, predictors were asked to provide quality estimates for protein structure models. The QA category uses the automatically produced models that are traditionally distributed to CASP participants as input for predictions. Predictors were asked to provide an index of the quality of these individual models (QM1) as well as an index for the expected correctness of each of their residues (QM2). We computed the correlation between the observed and predicted quality of the models and of the individual residues achieved by the participating groups and evaluated the statistical significance of the differences. We also compared the results with those obtained by a "naïve predictor" that assigns a quality score related to how close the model is to the structure of the most similar protein of known structure. The aims of a method for assessing the overall quality of a model can be twofold: selecting the best (or one of the best) model(s) among a set of plausible choices, or assigning a nonrelative quality value to an individual model. The applications of the two strategies are different, albeit equally important. Our assessment of the QA category demonstrates that methods for addressing the first task effectively do exist, while there is room for improvement as far as the second aspect is concerned. Notwithstanding the limited number of groups submitting predictions for residue-level accuracy, our data demonstrate that a respectable accuracy in this task can be achieved by methods relying on the comparison of different models for the same target.

Proteins 2007; 69(Suppl 8):175–183.
© 2007 Wiley-Liss, Inc.

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

INTRODUCTION

The quality of a protein structure model dictates its correct usage.[1] Models of moderate quality can be used as frameworks for rationalizing the results of experiments or evaluating the likelihood of a predicted protein function, but are of limited use for applications such as inhibitor design or docking. For comparative models, knowledge-based rules do exist for assessing beforehand the expected quality of a model on the basis of the sequence similarity between the target protein and the homologous template[2] or of the pair-wise similarity between the elements of a multiple sequence alignment.[3] Yet, the advent of novel techniques for alignments and for combining different templates might very well produce models with accuracy higher than that expected on the basis of sequence comparisons alone. Furthermore, fold recognition and fragment-based methods have matured and can produce models of respectable quality even in the absence of detectable sequence similarity with proteins of known structure.[4–6] In these cases, sequence similarity cannot be exploited for estimating in advance the expected quality of a model.

Methods able to predict the quality of a model on the basis of its coordinates alone would therefore be of high

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat
The authors state no conflict of interest.
Grant sponsor: European Commission; Grant number: LSHG-CT-2003-503265; Grant sponsor: NIH; Grant number: LM07085-01; Grant sponsor: MIUR grant LIBI—Laboratorio Internazionale di BioInformatica

*Correspondence to: Anna Tramontano, Department of Biochemical Sciences "Rossi Fanelli," University of Rome "La Sapienza", P.le Aldo Moro, 5, 00185 Rome, Italy. E-mail: anna.tramontano@uniroma1.it

Received 6 April 2007; Revised 8 June 2007; Accepted 12 June 2007
Published online 6 August 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21669

value for the end users of a model. At the same time they would contribute to the improvement of fold recognition and fragment-based methods as well as meta predictor strategies. In all these cases, the evaluation of several alternative models and the selection of the most likely one is an essential and crucial step.

This is the context within which CASP organizers decided to introduce a new prediction category, model quality assessment, whose evaluation is described here. To conduct the model quality prediction experiment, the organizers took advantage of the fact that models produced by servers participating in CASP were made available to all predictors in the course of the experiment, soon after the target release. The public distribution of server results served the purpose of avoiding overload on the servers and the released models traditionally were used by some human-expert groups as a starting point for their advanced structural modelling work. In CASP7, the server models were used also as targets for model quality prediction. Participating groups were asked to submit estimates for the quality of these models before the corresponding experimental structure was available.

Predictors were given the opportunity of submitting quality predictions for the model structure as a whole (Model Quality Mode 1, QM1, one value per model) and/or on a residue-by-residue basis (Model Quality Mode 2, QM2). At the end of the experiment, the observed quality of the server models (according to the results of the CASP automatic evaluation) was compared with the values submitted by the quality predictors.

In the following, we describe the set-up of the experiment and its results. We also discuss the limitations of the present methods and indicate where, in our opinion, efforts should be focused in the future.

MATERIALS AND METHODS

The target models used for quality prediction are accessible at http://predictioncenter.org/casp7/SERVER_HTML/tarballs/

The submitted quality assessment predictions are available at http://predictioncenter.org/download_area/CASP7/predictions/QA283-386.tar.gz

Some numerical evaluation data for the submitted quality predictions can be found at http://predictioncenter.org/casp/casp7/public/cgi-bin/qa_summary.cgi

All statistical analyses were performed using in-house scripts based on the R environment.[7]

We devised a "naïve predictor" called BLAST/LGA. This method identifies the protein of known structure closest to the target protein by running a PSI-BLAST[8] search with default parameters on the nr database[9] frozen at the time of release of the target (maximum number of iterations = 5). Next, it superimposes each model of the target to the identified template using LGA.[10] The quality prediction is simply the LGA_S score divided by 100 (as to normalize the result in the 0.0–1.0 range).

Correspondence between numerical group IDs used in this article and group names can be found at http://predictioncenter.org/casp7/meeting_docs/groups_by_number.pdf.

RESULTS

Description of the experiment

23,864 models for 95 targets were submitted by automatic servers and made available to predictors via the CASP7 web site (http://predictioncenter.org/casp7). Twenty-eight groups participated in the QM1 experiment, nine of which also submitted QM2 predictions.

In QM1, participants were asked to submit a quality score comprised between 0.0 and 1.0 for each model. The quality was meant to be on an absolute scale and directly correlated with the quality of the model, i.e., 0.0 score had to designate a very poor model while 1.0 was meant to identify a model perfectly coinciding with the target. However, some predictors submitted relative values for each target, i.e., they assigned a quality value of 1.0 to the predicted best model for each target, independently on how good such best model was. As we show later, our assessment of the predictions took both possibilities into account.

QM2 predictors were asked to assign an error estimate (in Å) to each residue of each model, or a null value indicated by "*X*" if they chose not to submit an estimate for one or more residues.

Quality mode 1: Overall quality of a model

Two groups (105 and 694) submitted predictions for only one target and were not considered further in the QM1 analysis. To assess the degree of success of the remaining participants, we computed the Pearson's correlation coefficient between the predicted quality and the observed GDT-TS value for each model computed and made available by the Prediction Center (see Methods Section). We computed the Pearson's coefficient both on a target by target basis (to take into account cases where predictors had normalized their quality assignment in the range 0.0–1.0 for each target) and by considering all models for all targets together to assess whether the predicted quality was indicative of the effective accuracy of each model, independently on the specific target.

For scoring the predictions we explored different routes, which are schematically illustrated in Figure 1. The results discussed here were obtained by using the estimated Pearson's correlation coefficient *r* between the variables of interest, i.e., the predicted accuracy score and the observed GDT-TS values. However, the final ranking of the groups is essentially unaffected by the choice of the scoring scheme (data not shown).

DOI 10.1002/prot

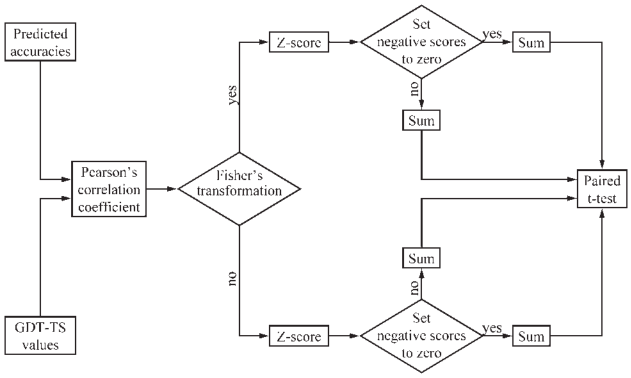

Figure 1

Outline of the methodology used for assessing the quality prediction category mode 1.

We calculated the distribution of r for each target, its average and standard deviation, and assigned a Z-score to each of the predictions. The distribution of the r values is only approximately normal, as verified by both visual inspection and Shapiro–Wilk test,[11] hence the Z-scores do not have a theoretically correct statistical significance. Nevertheless, we think that the Z-score is still appropriate as scoring system in this context, since it takes into account the difficulty of predicting the quality of the models for each target. Therefore, the sum over all targets of the Z-scores of a method is a reasonable estimate of its overall performance.

The sum of the Z-scores over all targets for each predictor is shown in Figure 2. Negative Z-scores were set to zero in order not to penalize more innovative and hence riskier methods. The significance of the differences among the scores for different groups was assessed by a paired t-test on the common set of predicted models (Table I and Supplementary Table I).

It is apparent that two groups outperform the others, namely groups 634 (Pcons) and 556 (Lee). The difference between their predictions and those obtained by the remaining groups is statistically significant as shown by the results of the paired t-test.

The procedure described above was repeated by dividing the targets into their respective prediction categories: Template-based (TBM) and Template-Free modelling (FM). We faced the problem that the category assignment is domain-based while quality predictions are submitted for whole targets. For this reason, the category-based analysis was limited to single domain proteins (68) and to proteins whose domains are assigned to the same prediction category (22). These 90 proteins add up to about 95% of the total number of assessed targets. For the purpose of this analysis, domain targets assigned to both the TBM and FM categories[12] were considered part of the TBM category. The results are shown in Figure 3.

Not much changes in terms of results when targets are clustered according to their respective category. Group 713 (Circle-QA) stands out among the other groups for FM models, although it should be mentioned that there are only 10 FM targets and is therefore unclear whether this conclusion can be generalized.

Next, we computed the global correlation coefficient for all the models of all targets taken together (Fig. 4). In this case, the overall correlation coefficient indicates the ability of the methods to assign nonrelative scores. The statistical significance of the difference between the correlation coefficients achieved by two different groups was computed as follows:

Given two Pearson's correlation coefficients r_1 and r_2, to test the null hypothesis that they derive from the same distribution, we applied the Fisher's transformation[13]

$$z' = (\ln(1+r) - \ln(1-r)) \times 0.5$$

The variable $z_1' - z_2'$ 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 associated P value indicates the probability that the two predictions with correlation coefficient r_1 and r_2 are statistically indistinguishable.

As shown in Figure 4, only group 556 is clearly and significantly better than the remaining ones (Table II and Supplementary Table II).

After the experiment, we were made aware that a program bug had introduced errors in some of the predictions submitted by group 634. Our assessment does not take this into account.

There is one more evaluation parameter that we thought could be relevant from the point of view of a user, namely how much would a user lose, should he or she select the model ranked as first by each method. In

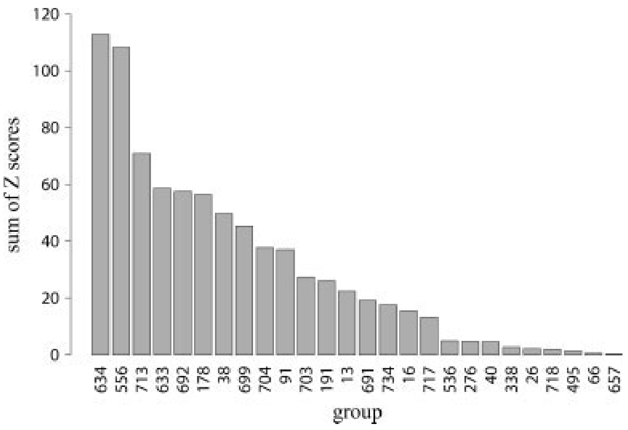

Figure 2

Overall scores of the groups participating in the QM1 experiment.

Table I
Statistical Comparisons Among the Top 16 Groups Submitting QM1 Predictions on a Target-by-Target Basis

	634	556	713	633	692	178	38	699	704	91	703	191	13	691	734	16
634		93 22,905	95 23,002	95 22,980	95 22,731	95 17,758	85 12,030	95 22,002	95 23,146	94 19,927	69 16,696	61 14,634	78 19,149	95 22,573	92 20,136	86 9282
556	0.000		93 22,901	93 22,514	93 22,265	93 17,355	84 11,917	93 21,912	93 22,797	92 19,837	68 16,544	61 14,728	76 18,986	93 22,175	91 20,032	85 9328
713	−0.080	−0.080		95 22,607	95 22,362	95 17,566	85 12,001	95 22,032	95 22,902	94 20,011	69 16,463	61 14,539	78 19,140	95 22,259	92 19,957	86 9296
633	0.124	−0.112	0.038		95 22,730	95 17,658	85 12,028	95 21,594	95 22,964	94 19,519	69 16,545	61 14,588	78 18,823	95 22,549	92 20,019	86 9147
692	−0.133	−0.125	0.036	−0.001		95 17,520	85 12,028	95 21,356	95 22,715	94 19,300	69 16,300	61 14,369	78 18,589	95 22,336	92 19,838	86 9041
178	0.160	0.117	0.062	0.033	0.029		85 12,035	95 17,328	95 17,824	94 16,752	69 13,160	61 10,742	78 14,552	95 17,578	92 16,363	86 6932
38	0.114	0.095	0.022	0.003	0.002	0.019		85 11,686	85 12,053	84 11,874	60 8330	60 8618	72 10,302	85 11,973	85 11,760	82 5165
699	−0.168	−0.168	0.120	−0.063	−0.051	−0.016	−0.044		95 21,883	94 19,331	69 15,940	61 13,608	78 18,286	95 21,277	92 19,129	86 8894
704	−0.185	−0.175	0.096	−0.059	−0.055	−0.020	−0.053	−0.011		94 19,743	69 16,761	61 14,719	78 19,066	95 22,611	92 20,185	86 9246
91	−0.164	−0.158	−0.083	−0.045	−0.034	−0.040	−0.081	−0.005	−0.006		69 14,438	60 12,614	77 16,545	94 19,246	91 18,339	85 8086
703	−0.193	−0.170	0.111	−0.066	−0.057	−0.034	−0.118	0.040	−0.001	0.039		40 9461	53 12,875	69 16,303	67 14,671	61 6518
191	0.146	0.159	0.074	0.051	0.037	−0.044	0.018	−0.036	−0.025	−0.019	−0.019		53 12,716	61 14,309	61 13,378	61 6449
13	0.212	0.213	0.121	0.081	0.071	0.025	0.039	0.021	0.020	0.017	0.009	0.037		78 18,511	77 17,103	72 7855
691	−0.331	−0.321	0.245	−0.207	0.195	−0.182	−0.212	0.194	0.144	0.140	0.150	−0.173	−0.117		92 19,758	86 8995
734	−0.300	−0.290	−0.214	−0.173	−0.161	−0.147	−0.132	−0.142	−0.120	0.122	−0.105	−0.110	−0.102	0.023		86 8252
16	0.261	0.256	0.167	0.122	0.109	0.167	0.204	0.102	0.067	0.090	0.044	0.088	−0.039	−0.100	−0.008	

Results of the paired t test. Pearson's coefficients for targets of common models were computed, and their distributions compared. The upper half of the matrix reports the numbers of common targets and models between the two groups corresponding to the row and the column. Differences in the means are shown in the lower left cells, which are shaded when their value is not statistically significant (i.e. the corresponding *P* value is greater than 0.01).

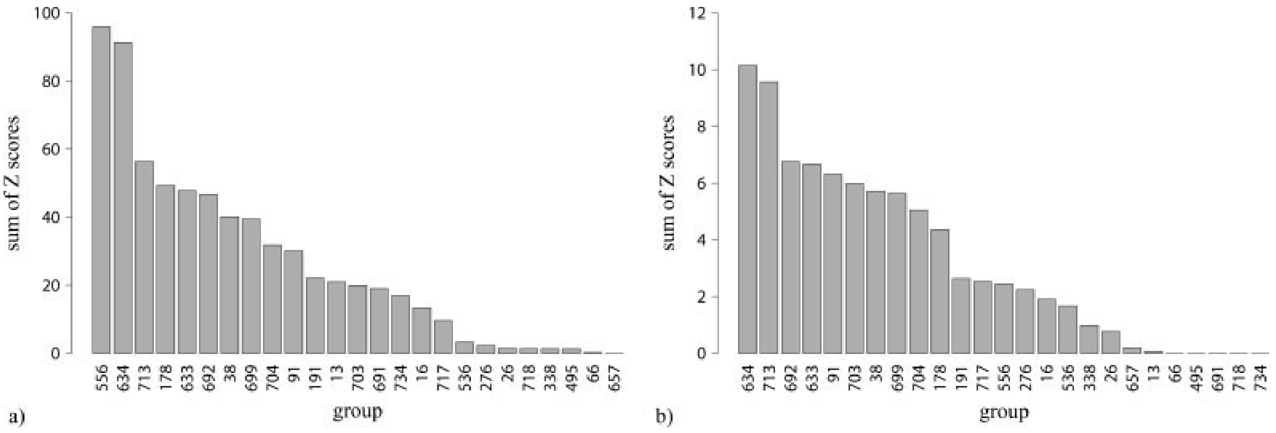

Figure 3

*Scores of the groups participating in the QM1 experiment for TBM (**a**) and FM (**b**) targets. Scores are computed on a per-target basis.*

other words, what is the GDT-TS difference (ΔGDT) between the model ranked as first by a method and the best model for each target? The results are rather different for different targets (Fig. 5); however, the average ΔGDT over all targets ranges between 3.18 and 27.74. The average ΔGDT values for groups 556 and 634 are 6.77 and 7.28, respectively. The ΔGDT of groups 556 and 634 are lower than the average values for 72 and 65 targets, respectively, and this difference is statistically significant ($P < 0.01$).

Whether or not an average improvement of a few units of GDT-TS is significant, obviously depends upon the specific application of the model. We have previously shown[14] that minor improvements in the quality of a model can make a substantial difference for some applications. In others, such as using the model as a framework for designing experiments, such a difference is probably not crucial unless it detects improvements in functionally important regions of the proteins. Visual inspection of the differences between the best models and those predicted as best by the top performing groups does not seem to indicate that this is the case (data not shown).

The average ranks of the best model in the scale provided by groups 634 and 556 are around positions 37 and 50, respectively. This implies that, although these methods can provide a ranking significantly correlated with the experimental one, they are not able to consistently select the best model. This somehow might limit the possibility of using these scoring methods as a target function for refinement and suggests that more effort should be put in this direction.

Quality mode 2: Residue-based quality prediction

The evaluation of the QM2 predictions was based again on the Pearson's correlation coefficient. We computed the sum of *Z*-scores as described above (Fig. 6). The statistical significance of the differences among these scores was assessed through a paired t test (Table III) using the common residues of the common models of the common targets. We discarded from the analysis 511 models (about 2.2% of the total number of models) for which less than four predictions were submitted.

Group 556 did not participate in the QM2 experiment. Group 634 still performed significantly better than the others (Fig. 6 and Table III). Two examples of predictions of this group are shown in Figure 7, where the predicted and observed errors in model coordinates are shown. The two examples are for targets T0307 and T0346, which have very different prediction difficulties. The first is among the hardest-free modelling targets, while the second is a rather straightforward comparative modelling

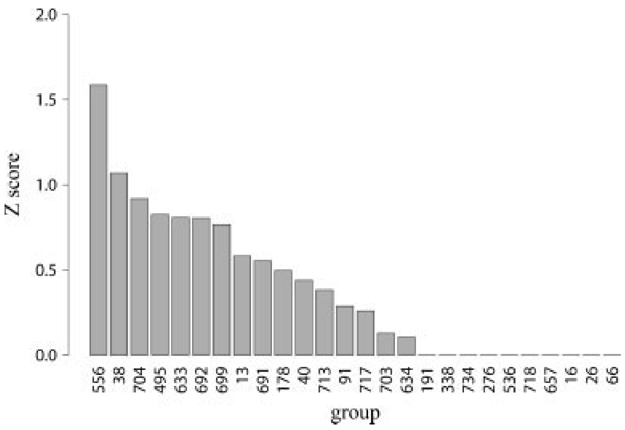

Figure 4

Scores of the predictions of the groups participating in the QM1 category when all quality predictions are pooled together.

Table II

Statistical Comparisons Among the QM1 Predictions Submitted by the Top 13 Groups when Pooling All Targets Together

	556	38	704	495	633	692	699	13	691	178	40	713	91
556	23,281												
38	34.886	12,057											
704	−50.586	−6.911	23,254										
495	8.496	−1.997	0.722	281									
633	−55.985	−11.493	5.558	−0.135	22,962								
692	−55.953	−11.572	5.664	−0.154	−0.121	22,713							
699	−57.583	−13.064	7.479	−0.439	−1.951	−1.826	22,385						
13	63.312	19.276	15.013	1.693	9.662	9.522	7.735	19,474					
691	−66.946	−20.715	16.709	−1.864	−11.137	10.987	9.123	−1.045	22,615				
178	65.001	21.663	17.869	2.222	12.640	12.497	10.744	−3.091	2.179	17.832			
40	23.156	−9.364	6.897	2.346	5.103	5.062	4.465	−1.833	1.487	−0.726	1270		
713	−74.317	−26.454	−23.653	−2.909	−18.012	−17.841	−15.937	−7.548	−6.755	−4.142	−0.696	23.385	
91	−74.985	−28.617	−26.180	−3.447	−20.741	−20.570	−18.727	−10.545	−9.884	−7.187	−1.821	−3.400	20.235

Results of the *Z* test. Pearson's coefficients for all predictors were computed, and their corresponding Fisher's transformations compared. Diagonal entries report the number of models assessed by the corresponding group. *Z* values are shown in the lower half of the matrix. Cells are shaded when the difference is not statistically significant (i.e. the corresponding *P* value is greater than 0.01).

target.[15] In both cases the Pearson's correlation coefficient is rather high: 0.83 and 0.98, respectively.

We also looked at the performance of the nine groups on the TBM and FM categories separately. For this purpose, we analyzed all target domains, except T0320_2 which was described as decoration in the CASP7 domain classification. All domains defined as TBM/FM (T0304, T0321_1, T0348, and T0382) were included in the TBM category. Results of this test are illustrated in Figure 8. Group 634 outperforms all other participants in both cases.

Naïve predictor

We decided to verify whether the results achieved in the experiment could have been reproduced, or even outperformed, by simple methods based on our experience in previous CASPs. Indeed it has been apparent over the years that in the TBM category it is very difficult to consistently obtain a model that is better than the template.[15–19] We thus designed a naïve predictor that scores TBM models by assigning a value related to their distance from the closest template. Models for multi-domain targets were not considered.

The method (named BLAST/LGA) assigns a quality related to the similarity of the model with the best template measured using LGA. It selects as template the first sequence of known structure found in a PSI-BLAST search against the nr database frozen at the release date of each target. It was applied to the 45 targets for which PSI-BLAST identified a

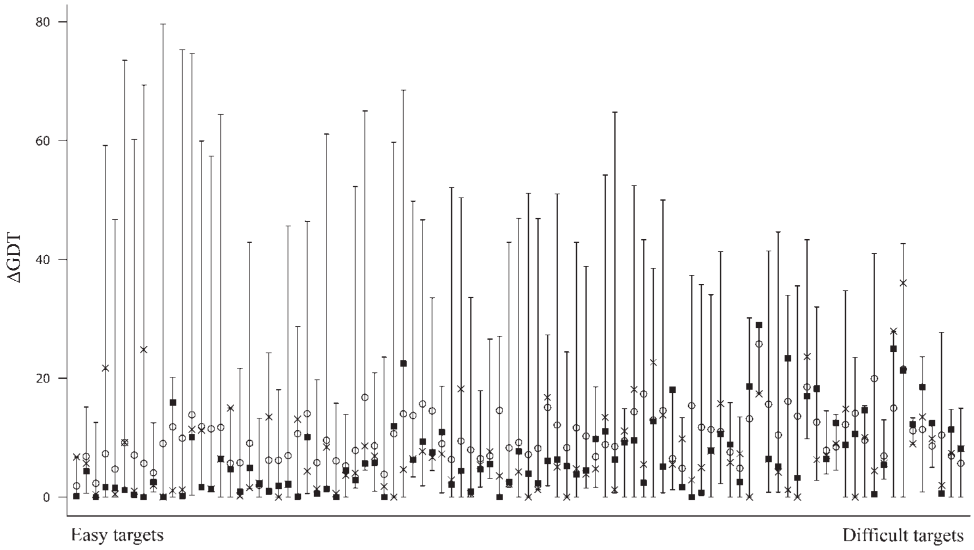

Figure 5

ΔGDT values for predictors 556 (black squares) and 634 (crosses) compared with the average ΔGDT for all groups (empty circles) assessed in the QM1 experiment. Vertical lines indicate the maximum and minimum average ΔGDT values for each target.

DOI 10.1002/prot

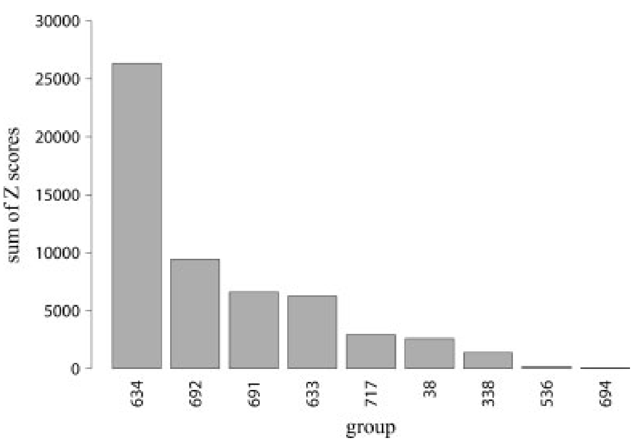

Figure 6
Scores of the nine groups participating in the QM2 experiment.

structural template (11,588 models). In 19 cases, the first hit of PSI-BLAST is also the best structural template.

We compared the results obtained by the naïve predictor with those of all other methods in the TBM and TBM/FM categories. Its results were not used to compute the average and standard deviation of the correlation coefficient distributions to avoid biasing the results. In other words, we computed the *Z*-score of the predictions obtained by BLAST/LGA using the average and standard deviation values of the distribution of all submitted predictions.

Interestingly, for TBM models, only 556 and 634 clearly outperform the naïve predictor, and only a few other methods achieve results of comparable quality. (Fig. 9 and Supplementary Table III).

Strategies and methodology of the best performing groups

The methodology used by group 634 (Pcons) is illustrated in a article in this same issue.[20] The strategy adopted by group 556 (Lee), albeit very successful, is of little use outside the CASP experiment. This group produced very good models for most targets and, subsequently, compared all target models with their own predictions, assigning a value related to the distance of the analyzed model from their own.

The methodology of group 634 is a consensus approach based on the multiple predictions submitted by several groups, and therefore is suitable for selecting good models among a set of alternative ones, but it cannot be used for assessing the quality of a model on its own.

Interestingly, the strategy employed by group 713 (Circle-QA) is solely based on the coordinates of the analyzed model. This group did not perform better than groups 556 or 634 and it did not do much better than our naïve predictor in QM1. Nevertheless we believe it will be interesting to follow its developments in the future. We would like to encourage the development of quality assessment programs capable of reliable estimation of model quality based on structural features of the model alone and not depending upon the quality of other models available in the decoy set.

DISCUSSION

It is our opinion that assessment of quality prediction methods is of outmost importance for the protein structure prediction field. An effective method for discriminating between models of reasonable and yet different quality would have a tremendous impact on several

Table III
Statistical Comparisons Among Groups Participating in QM2

	634	692	691	633	717	38	338	536	694
634		22,625 4,213,326	22,469 4,207,836	22,873 4,267,995	20,728 3,828,764	11,983 2,268,833	22,530 4,170,223	728 95,322	224 23,244
692	−0.196		22,233 4,155,477	22,624 4,213,215	20,323 3,746,851	11,981 2,268,389	22,102 4,083,631	728 95,322	224 23,244
691	−0.282	0.087		22,445 4,201,509	20,221 3,761,522	11,926 2,265,890	21,994 4,097,904	724 94,944	221 22,920
633	0.259	0.063	−0.023		20,561 3,799,250	11,981 2,268,588	22,348 4,137,680	728 95,322	223 23,133
717	−0.472	−0.275	−0.189	−0.209		10,944 2,067,188	20,783 3,844,491	702 93,460	219 22,676
38	0.275	0.100	0.025	0.033	−0.202		11,951 2,256,219	0 0	113 12,535
338	0.481	0.285	0.200	0.221	0.006	0.209		740 97,679	224 23,233
536	0.491	0.224	0.071	0.188	−0.001	NA	−0.026		0 0
694	−0.181	0.000	0.090	0.007	−0.160	0.001	0.126	NA	

Results of the paired *t* test. Pearson's coefficients for residues of common models were computed, and their distributions compared. The upper half of the matrix reports the numbers of common models and residues between the groups in the column and the row. Differences in means are shown in the lower half of the matrix. Cells are shaded when their value is not statistically significant (i.e. the corresponding *P* value is greater than 0.01). "NA": no common predictions between the two groups.

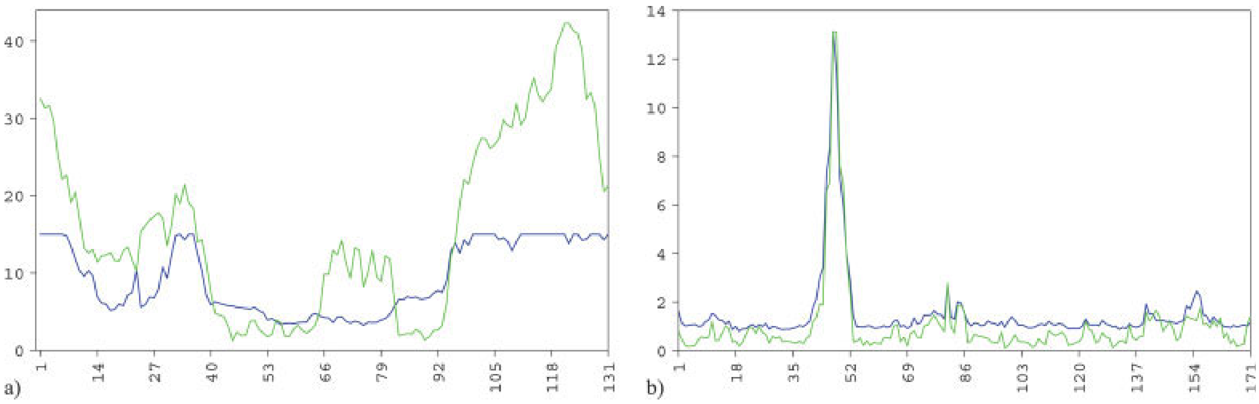

Figure 7

*Plots of the predicted (blue lines) and observed (green lines) Cα distances between target and models T0307TS464_1 and T0346TS414_1 in (**a**) and (**b**), respectively. Both predictions were submitted by group 634.*

aspects of protein structure prediction. Methods able to correctly select the best model among a set of alternative ones are useful in optimization techniques, in fold recognition and as selector in meta predictions. Our results show that Pcons is the best method among those participating in the experiment and that it can be reliably used for selecting models that are, on average, about 7 units of GDT-TS from the best available one.

Although useful, however, the strategy adopted by Pcons and by similar tools cannot be used to assess the quality of a single model. A method able to assign nonrelative scores would be a very interesting tool for the biological community. It would allow the quality of automatically produced models or of models present in different model databases[21–23] to be judged, and therefore result in a more effective use of the resources.

Among the methods able to assign a nonrelative score to a model, only group 556 showed a reasonable level of accuracy, while others did not reach the same performance and could be often outperformed by very simple strategies such as the one used by our naïve predictor simply based upon the distance between the model and the best available template. The "success" of the latter has also another implication: among the comparative models produced by automatic servers, the best ones seem to be those that deviate less from the template. This conclusion has to be taken with care for several reasons. First, it is limited to single domain comparative modelling targets, and, second, it only concerns automatic models produced by publicly available servers. The topic of the extent by which modelling methods improve over the single best template is also discussed in the Tem-

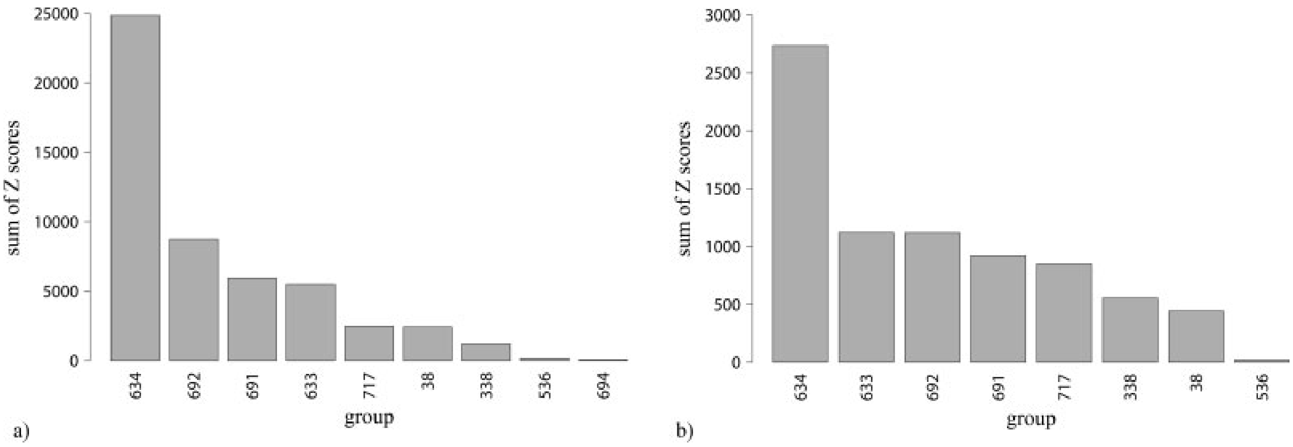

Figure 8

Scores of the groups participating in the QM2 experiment for targets in the TBM and FM categories (panels a and b, respectively).

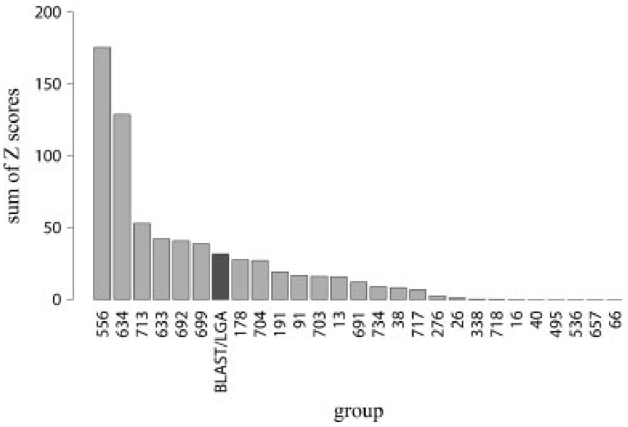

Figure 9

Comparison of the performance of the naïve method BLAST/LGA with all other methods submitting QM1 predictions for TBM targets.

plate Based assessment article[24] and in the article discussing progress in different CASP experiments[15] in this issue. The unifying theme of the results of the different analyses is that methods able to obtain models better than the best available template are not the majority and that the improvement is rarely very substantial. The results presented here are well in line with these conclusions.

Other tools that could be very valuable for the biological community are, obviously, those for assigning the reliability of individual residues in a model, i.e., those assessed in the QM2 category. Effective methods could highlight reliable regions of a model, and increase its usefulness even when the overall quality is not the highest.[25] The results of this aspect of the experiment are not exceptionally encouraging in our view. The strength of the CASP assessment relies on a statistically sound analysis of the results and too few groups participated in QM2 to allow us to derive reliable conclusions. Only the Pcons method seems to be able to identify the best-predicted regions in a reasonable number of cases, while the reliability of other methods lags behind.

In several occasions, the establishment of a new category in CASP has prompted more and more scientists to tackle the corresponding problem. We hope this is the case for quality prediction too and look forward to a larger participation in the next round.

REFERENCES

1. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294:93–96.
2. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5:823–826.
3. Cozzetto D, Tramontano A. Relationship between multiple sequence alignments and quality of protein comparative models. Proteins 2005;58:151–157.
4. Wang G, Jin Y, Dunbrack RL, Jr. Assessment of fold recognition predictions in CASP6. Proteins 2005;61(Suppl 7):46–66.
5. Vincent JJ, Tai CH, Sathyanarayana BK, Lee B. Assessment of CASP6 predictions for new and nearly new fold targets. Proteins 2005;61(Suppl 7):67–83.
6. Jauch R, Yeo HC, Kolatkar PR, Clarke ND. Assessment of CASP7 structure predictions for template free targets. Proteins 2007;69 (Suppl 8):57–67.
7. R Development Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2006.
8. 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.
9. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2007;35(Database issue):D5–D12.
10. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
11. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika 1965;52:591–611.
12. Tress M, Clarke N, Ezkurdia I, Kopp J, Read R, Schwede T. Domain definition and target classification for CASP7. Proteins 2007;69 (Suppl 8):10–18.
13. Anderson TW. An introduction to multivariate statistical analysis. Hoboken, New Jersey: Wiley-Interscience; 2003.
14. Giorgetti A, Raimondo D, Miele AE, Tramontano A. Evaluating the usefulness of protein structure models for molecular replacement. Bioinformatics 2005;21(Suppl 2):ii72–ii76.
15. Kryshtafovych A, Krzysztof F, Moult J. Progress from CASP6 to CASP7. Proteins 2007;69(Suppl 8):194–207.
16. Tramontano A, Leplae R, Morea V. Analysis and assessment of comparative modeling predictions in CASP4. Proteins 2001; 45(Suppl 5):22–38.
17. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. Proteins 2003;53(Suppl 6):352–368.
18. 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.
19. Kryshtafovych A, Venclovas C, Fidelis K, Moult J. Progress over the first decade of CASP experiments. Proteins 2005;61(Suppl 7):225–236.
20. Wallner B, Elofsson A. Prediction of global and local model quality in CASP7 using Pcons and ProQ. Proteins 2007;69(Suppl 8):184–193.
21. Kopp J, Schwede T. The SWISS-MODEL repository: new features and functionalities. Nucleic Acids Res 2006;34(Database issue): D315–D318.
22. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D, Shen MY, Kelly L, Melo F, Sali A. MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 2006;34(Database issue):D291–D295.
23. Castrignano T, De Meo PD, Cozzetto D, Talamo IG, Tramontano A. The PMDB protein model database. Nucleic Acids Res 2006; 34(Database issue):D306–D309.
24. Kopp J, Bordoli L, Battey J, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. Proteins 2007;69(Suppl 8):38–56.
25. Pizzi E, Tramontano A, Tomei L, La Monica N, Failla C, Sardana M, Wood T, De Francesco R. Molecular model of the specificity pocket of the hepatitis C virus protease: implications for substrate recognition. Proc Natl Acad Sci USA 1994;91:888–892.