PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

NEW FOLDS: ASSESSMENT

Assessment of CASP8 structure predictions for template free targets

Moshe Ben-David,[1] Orly Noivirt-Brik,[1] Aviv Paz,[1] Jaime Prilusky,[2] Joel L. Sussman,[1,3] and Yaakov Levy[1]*

[1] Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel
[2] Bioinformatics Unit, Weizmann Institute of Science, Rehovot 76100, Israel
[3] The Israel Structural Proteomics Center, Weizmann Institute of Science, Rehovot 76100, Israel

ABSTRACT

The biennial CASP experiment is a crucial way to evaluate, in an unbiased way, the progress in predicting novel 3D protein structures. In this article, we assess the quality of prediction of template free models, that is, *ab initio* prediction of 3D structures of proteins based solely on the amino acid sequences, that is, proteins that did not have significant sequence identity to any protein in the Protein Data Bank. There were 13 targets in this category and 102 groups submitted predictions. Analysis was based on the GDT_TS analysis, which has been used in previous CASP experiments, together with a newly developed method, the OK_Rank, as well as by visual inspection. There is no doubt that in recent years many obstacles have been removed on the long and elusive way to deciphering the protein-folding problem. Out of the 13 targets, six were predicted well by a number of groups. On the other hand, it must be stressed that for four targets, none of the models were judged to be satisfactory. Thus, for template free model prediction, as evaluated in this CASP, successes have been achieved for most targets; however, a great deal of research is still required, both in improving the existing methods and in development of new approaches.

Proteins 2009; 77(Suppl 9):50–65.

Key words: structure prediction; free modeling; Q measure; CASP.

INTRODUCTION

The biennial CASP experiment is a crucial way to evaluate, in an unbiased way, the progress in predicting novel 3D protein structures. This is the eighth such experiment which have taken place at 2-year intervals starting in 1994.[1,2] These experiments are done in a "double-blind" manner, that is, the predictors only have access to the amino acid sequences of the proteins to predict and not to the 3D structures of the targets, and the assessors only know the groups by "group numbers" and the actual scientists associated with each group are not known during the assessment process.

There has been significant progress in the novel structure prediction since the first CASP experiments, which is based largely on biased sampling of structural fragments from the PDB as a way to assemble initial models, an idea that is more than 24 years old,[3–6] as was discussed in CASP7.[7] However, protein structure prediction is still a very challenging problem, and an objective way to assess it is also much more difficult than commonly thought. As Jauch *et al.*[7] wrote: "In assessing structure prediction, it is useful to have quantitative metrics that can identify objectively the models that are most similar to the target structure. However, it is not a simple matter to define such metrics. It is even problematic to define what one means by structural similarity. Indeed, any definition of structural similarity,

Additional Supporting Information may be found in the online version of this article.
The authors state no conflict of interest.
Grant sponsor: Kimmelman Center for Macromolecular Assemblies, Erwin Pearl, the Divadol Foundation, the Nalvyco Foundation, the Bruce Rosen Foundation, the Jean and Julia Goldwurm Memorial Foundation, the Neuman Foundation, the Kalman and Ida Wolens Foundation, Center for Complexity Science.
*Correspondence to: Yaakov Levy; Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel. E-mail: koby.levy@weizmann.ac.il.
Received 3 May 2009; Revised 4 August 2009; Accepted 7 August 2009
Published online 21 August 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22591

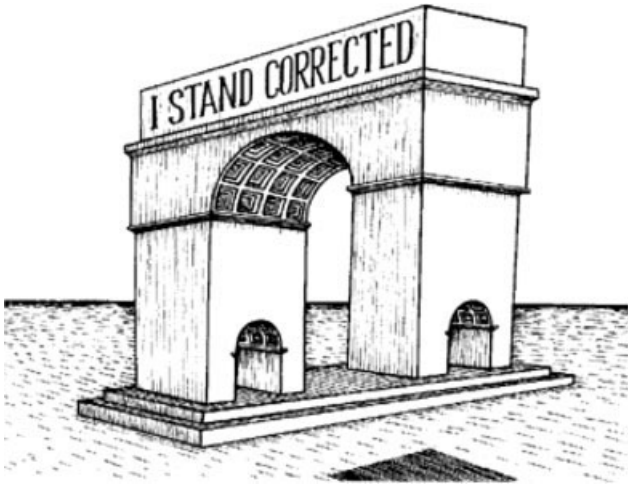

Figure 1
What you see is what you want to see.

and any quantitative measure of similarity, is an implicit (and imperfect) statement about what is considered to be important in a structure prediction."

GDT_TS[8] is a widely used measure of backbone similarity for evaluating template-based models and has been used over the last several CASP experiments.[7] In parallel, GDT_TS has been used to assess new fold predictions; however these are particularly difficult to objectively assess, as even for single protein target domains as small as ~100 amino acids, often few of the models have an RMS deviation under 10 Å for Cα's. Thus it is not clear for this class of poorer models how well the GDT_TS scores correlate with what structural biologists would consider, via visual examination, to be a good model. Because of this, in previous CASP experiments, the assessors had to rely to a very large extent on visual inspection of the *ab initio* models to judge which ones were the best. A feeling for this kind of difficulty is illustrated in Figure 1. A number of methods were tested in previous CASP experiments, attempting to objectively and quantitatively assess the quality of predicted 3D structures, but so far none have proved to be more reliable than GDT_TS. In the current CASP experiment, we have developed the "*Q*" score, which is an objective way to compare a model to its experimentally determined structure without requiring any initial 3D superposition. For several targets, the best models indicated by the *Q* measures were correlated with those suggested by the GDT_TS score. Furthermore, the *Q* score enriched the list of candidates for best model, which were further investigated visually. Versions of the *Q* score proved to be useful in visualizing similarity between targets and their corresponding models and to provide a microscopic understanding of the successes and limitations of the predictions, which is not available using the GDT_TS score. The *Q* score, therefore, can quantify the accuracy of the predictions and can highlight regions or aspects that were well or poorly predicted, as well as quantifying global accuracy.

In CASP8, as contrasted with previous CASP experiments, only single domains were considered for template free predictions. The CASP administration divided the targets into individual domains if template availability or relative positioning varied between those domains.[9] This resulted in the CASP8 template-free targets being shorter in length than in previous CASP experiments. Comparing CASP8 versus CASP7 and CASP6, the average lengths are 90.5, 106.2, 142.9 amino acids (see Fig. 2). This in turn makes it difficult to assess if, in fact, there is any improvement in the prediction of CASP vs. previous CASP experiments.

With the advent of large-scale structural genomics and structural proteomics initiatives,[10] many more structures are being determined with sequence identities less than 30% to known structures in the PDB,[11] and in fact, out of the 13 targets in the template free category, all came from structural genomics centres. However, out of these 13 targets, only two can be classified as actually new folds,[12] that is, T0397-D1 and T0496-D1. Therefore, these template-free assessments must make do with fewer examples than we might have wished.

METHODS

Q scores

One common limitation of measures that compare protein structures is the need to perform structural alignment. When the two structures are structurally aligned a quantitative comparison of their structures can be obtained. Although estimating the structural similarity by aligning the two structures is very common (e.g., using

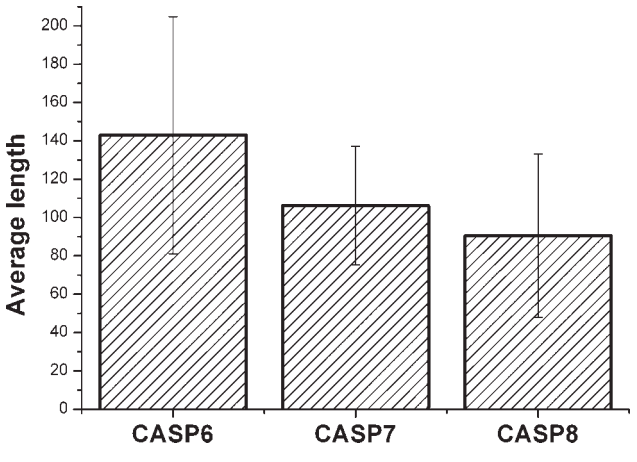

Figure 2
Comparison of average target lengths for FM models in CASP6, CASP7, and CASP8.

RMSD measure), the alignment can introduce large deviations due to a small perturbation (e.g., from a hinge in the structure) and suggest incorrectly that the two structures are different. This drawback of structural similarity measure based on structural alignment is addressed in the GDT_TS measure[8] by taking into account both local and global structure superpositions (more specifically, the GDT_TS measures the percent of residues from structure A that can be superimposed with structure B under several distance cutoffs, which are then averaged). Although the GDT score was proven useful in previous CASP experiments for selecting the models to be examined by visual inspection, it occasionally misses good candidates and does not provide a detailed molecular understanding of the quality of the prediction.

To evaluate the CASP8 predictions in detail and to highlight the origin of successes or failures of the predictions, we developed the Q score. It estimates the structural similarity between two given protein structures based on comparing their internal distances (thus overcoming the need for structural alignment). Our Q score is inspired by the Q measure developed by the Wolynes group for constructing the energy landscape of protein folding and for comparing structural complementarity of two structures.[13,14] To calculate the Q score, internal distances are calculated between the Cα atom of each residue i and all $N-1$ other Cα atoms in the protein, obtaining a matrix $\{r_{ij}\}$ (with $N(N\text{-}1)/2$ non-zero terms). The matrix for the target is designated as $\{r^0_{ij}\}$. For each pair of residues ($i-j>0$), Q_{ij} is calculated as $Q_{ij}=\exp[-(r_{ij}-r^0_{ij})^2]$. For a good prediction, $|(r_{ij}-r^0_{ij})|=0$, and $Q_{ij}=1$. For a very poor prediction $|(r_{ij}-r^0_{ij})| >> 0$, and $Q_{ij}=0$. Accordingly, each internal pairwise distance is compared to the corresponding distance in the target and gets a raw Q score between 0 and 1. Averaging all the Q_{ij}, a Q_{total} ($=\langle Q_{ij}\rangle$) measure is obtained that indicates the overall quality of the prediction. The Q_{total} measure is similar to the S_{contact} measure used by Grishin and his coworkers in CASP5.[15] We note that while a Q_{total} of 1.0 corresponds to an exact match of the two structures, Q_{total} of 0.4 for single domain proteins often indicates a reasonable prediction with RMSD of ~ 6 Å.[16]

For a given model, the Q_{ij} were sorted from $Q_{ij}=1$ to $Q_{ij}=0$. Note that since Q_{ij} is calculated also for $i-j=1$, which corresponds to adjacent Cα-Cα distances that should all equal 3.8 Å, all predictions will have some Q_{ij} close to unity. An averaged Q_{ij}, $\langle Q_{ij}\rangle=\frac{1}{M}\sum_M Q_{ij}$, is calculated for each step in the ranked list of Q_{ij} where M increases from 1 to $N(N-1)/2$. Values of $\langle Q_{ij}\rangle$ can be plotted against the fraction of pairwise distances involved in the calculation [i.e., $2M/N(N-1)$]. The better the prediction, the longer $\langle Q_{ij}\rangle$ stays high and the larger Q_{total} is. For a perfect prediction, $\langle Q_{ij}\rangle$ equals to 1 for any fraction of pairwise distances. For quite poor predictions, $\langle Q_{ij}\rangle$ will have low values even for small M (i.e., when small numbers of pairwise distances are included), and Q_{total} will be close to zero (see Fig. 3).

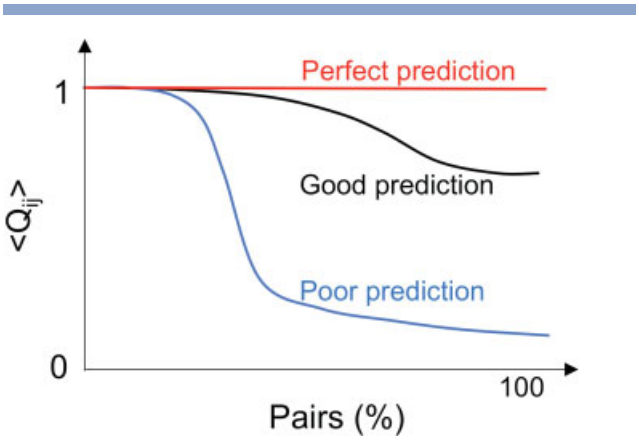

Figure 3

A schematic plot of the Q score along the fraction of pairwise distances involved in the Q calculations. The $\langle Q_{ij}\rangle$ is the normalized summation of Cα pairwise distance differences where the pairs are sorted based on their Q_{ij} (from 1 to 0). For a perfect prediction, $\langle Q_{ij}\rangle$ will be equal to 1 independently on the fraction of pairwise distances involved in its calculations. For a good prediction that includes some imperfect regions, $\langle Q_{ij}\rangle$ is expected to decrease when large number of pairs are involved, but Q_{total} (when all pairs are taken into account) will be still relatively large. For a poor prediction, $\langle Q_{ij}\rangle$ will be high only for low fraction of pairs and then will significantly decrease. Various features of these plots (the slope, the inflection point, and the Q_{total}) indicate the quality of the predicted structure. Such plots could be constructed when only subset of the pairwise distances are included such as inter-helical or inter-strands pairs or alternatively pairs that satisfy $|i\text{-}j|=20$ (Q_{short}) or $|i\text{-}j|>20$ (Q_{long}).

In the process of developing this final version of the Q measure, several variations of it were examined. We tried down-weighting the influence of long-range deviations with a relative error-Q measure where $Q_{ij}=\exp\left[-\left|\frac{r_{ij}-r^0_{ij}}{r_{ij}}\right|\right]$, this measure contains interesting information and although not used, it might be further considered in the future. A product-Q measure where $Q''_{ij}=\exp[-\frac{1}{M}\sum_M(r_{ij}-r^0_{ij})^2]$showed very high correlation with our original Q measure and, therefore, was not further considered.

To get structural information from the Q score we define two alternative measures: Q_{short} and Q_{long}, that are obtained by calculating Q for $|i-j|=20$ and for $|i-j|>20$, respectively. While Q indicates the overall quality of the model relative to the target, Q_{short} and Q_{long} indicate the quality of the secondary and tertiary structure of the prediction. Q_{short} of a given prediction will be calculated by averaging Q_{ij} when the best pair and 20, 40, 60, 80, and 100% of the ranked pairs that satisfy $|i-j|\leq 20$ are included. An averaged Q_{long} is similarly calculated. Obviously, correctly predicting interactions between residues far in the sequence is more challenging than predicting local interactions. High Q_{long}, therefore, indicates a good model and we found it to be correlated with the

GDT_TS score while Q_{short} was less correlated with GDT_TS. The Q score, in comparison to GDT_TS for example, can provide microscopic structural evaluation of the prediction by considering only subsets of the contact map. To indicate the packing and orientation of the secondary structure elements, we measure $Q_{\alpha\text{-helix}}$ and $Q_{\beta\text{-sheet}}$ by including only inter-helical or only inter-strand interactions, respectively, in the Q score. In the figures, we show Q_{short} and Q_{long} results for targets T0405-D1, T0482-D1, and T0510-D1. $Q_{\alpha\text{-helix}}$ and $Q_{\beta\text{-sheet}}$ are shown for targets T0482-D1, T0496-D1, and T0513-D2.

OK_rank

We combined Q_{short}, Q_{long}, GDT_TS,[8] and the MAMMOTH[17] Z-score into a score denominated OK_Rank. Namely, Q_{short}, Q_{long}, and GDT_TS scores were split into bins of one percent, and the models were ranked by their appropriate bin (i.e., two models with GDT_TS of 52.3 and 52.7 share the same GDT_TS rank). The MAMMOTH Z-score was used without any binning procedure (namely, the models with the top 15 "ranks" are the models with the top 15 scores). The OK_Rank score is obtained by the average of the four integer ranks. A table representing all models that were ranked in the top 15 bins of at least one of the scores was generated, and the assessors visually evaluated the models that were in the top 15 ranks of all four scores. Following this protocol, the number of candidate models for visual inspections was between 7 and 69.

Targets

CASP8 targets included thirteen free modeling (FM) targets, in which three targets were dedicated to server predictions and ten were classified as human/server targets. Three of the ten human/server targets were on the boundary between FM and template-based modeling (FM/TBM),[9] that is, T0405-D2, T0460-D1, T0476-D1.

Selection of models for visual assessment

For each target, the 20 best individual models according to GDT_TS scoring models, as well as the top scoring models according to the OK_Rank (39 models per target, on average) were visually inspected by three independent assessors (JLS, MB, and AP). Thus, some overlap in the targets assessed existed between the best GDT_TS scoring targets and the best OK_Rank targets. As long as a model from a certain group satisfied these conditions it was assessed, independently of the scores obtained by the other models from the same group, allowing the assessment of all five models from the same group. This is in contrast to previous CASP experiments (e.g., in CASP6) where only two models, at most, from the same group were permitted (i.e., the first model and the best GDT_TS scoring model).[18]

Visual inspection

Targets and models were visualized and aligned in a sequence dependent mode[8] by the SPICE DAS client.[19] More "challenging" targets were visualized and aligned in PyMOL,[20] which was subsequently used for the preparation of the figures. Each assessor independently chose the "best three models" for each target. As there were a few models from different groups that were identical, or almost identical, the assessors had the option of choosing more than three models as the "best three" (which was the case for almost all targets). On the other hand, for more challenging targets, less than three models were chosen due to the low quality of the models.

Scoring

To choose the best performing groups, the models selected for visual inspection were ranked by two different schemes, each scheme highlighting different aspects. Scoring Scheme A followed the strategy in which CASP is run; each group could submit five models for each target, and we wished to reward the groups that submitted more than one model that was considered by us as a top three model. A group was scored each time it appeared in the top three lists of each assessor, yielding a maximum score of 195 for all 13 targets and 150 for the 10 human/server targets (# of targets $\times$ # of models per target $\times$ # of assessors). As this scoring scheme does not necessarily provide data about the number of different targets each group has successfully modeled, we used, in parallel, an alternative scheme, that is, Scoring Scheme M, in which a group was counted once, irrespective of the number of times the assessors chose it for a specific target, to yield a maximum score of 13 for all targets and 10 for the human/server targets.

The best model for a given target was chosen on the basis of the agreement between the visual assessors on ranking a model as the #1 model in the top three lists; for eight targets all three assessors agreed unanimously on the specific best model. If the assessors did not reach a consensus on the best model, or could not choose any model for a particular target due to the low quality of all models, the best GDT_TS scoring model was designated for that target (see later, Targets: T0397-D1, T0443-D2, and T0461-D1). When multiple models too similar to be independent were the top choice, as seen in four of the targets, an attempt was made to identify a server model that could have acted as a template for the rest of the set, and that server or pair of servers was considered the best model.

The ranking of the best models as excellent, fair, and poor, as shown in Figure 4, was initially done subjectively, on the basis of visual inspection. It can be reproduced by a set of rules, however. Excellent best models are ones for which all assessors and scores agreed perfectly or very closely on the best, and for which GDT-TS

>50 and Dali-Z > 4. A best model is poor if any assessor judged there were no good models, or assessors and scores differed widely, and GDT-TS < 50 and Dali-Z <4. Fair best models have mixed scores and/or intermediate levels of agreement.

RESULTS

Results for individual targets

This section discusses the results of the 10 FM targets and three FM/TBM targets. Three of the 10 FM targets were designated as server only (S) predictions, whereas the other 10 were human/server (H/S) predictions. These 13 targets were assessed by visual inspection, in addition to the two main measures, GDT_TS score and OK_Rank. Each target is described briefly, and successful predictions or interesting observations are highlighted.

T0397-D1 (FM; H/S); PDB 3d4r

This domain contains a six-stranded, U-shaped antiparallel β-sheet that forms a 12-strand β-barrel in the biological-unit dimer. In addition, it is difficult to predict because of a very unusual topology with three crossover connections between strands. The assessors could not agree on any one model as being the best, and one of them judged that none of the models resembles the target. Many groups predicted the six antiparallel β-strands, but usually as a fairly flat sheet and never with the correct topology and arrangement to match that of the target. Interestingly, this is a clear case where the GDT-TS score prefers truly unacceptable models, fooled by a very approximate overlap of two strands on each side of the structure. There is very high similarity between the 20 top ranking GDT_TS models, which all predicted a flat sheet with a rather simple topology and a long α-helix between strands 3 and 4, whereas the target is strongly U-shaped with a very complex topology and a four-residue 3–10 helix (see Fig. 4). Of the other models, that got assessor votes and/or high scores on any of the quantitative measures, TS093_2 includes the greatest number of well-placed strands (Fig. 5) and TS020_5 the next most; however, even they are poor models.

T0405-D1 (FM; H/S)

This domain, which is a part of a larger protein structure, is relatively short and contains only three helices packed in a fold resembling an up-and-down three-helix bundle. The second helix is the longest and is bent, probably due contact with the other domain. Many groups had fairly good models for this target, especially for the second and third helices. For the first helix however, although secondary structure was predicted correctly, only a few groups could orient this helix the same as in the target structure. The top scoring model both in the GDT_TS (39.14) and OK_Rank, as well as in the visual assessment (ranked as best model by all three assessors), is from the Baker group (TS489_1). In this model, the three helices are nearly correctly positioned and oriented, but with minor imperfections in the connecting loops (Fig. 4). The second high scoring model is from GeneSilico (TS371_5). This model has a GDT_TS score of 36.68, it is second in the OK_Rank, and in the visual assessment it was chosen by all three assessors to be in the top three models. It is very similar to TS489_1, but the third helix is bent and oriented a bit differently than in the target, while it correctly predicts more of the loop regions (see Fig. 6). We note that the Q_{long} measure clearly indicates that models TS489_1 and TS371_5 are better than other models while the GDT_TS measure fails in classifying these two models as the two best ones.

T0405-D2 (FM; H/S)

This domain adopts an α + β-fold, composed of five α-helices and a six-stranded β-sheet, with anti-parallel topology, where strand six is broken by a sharp bend. The top scoring model with the highest ranking in both the GDT_TS and the OK_Rank is from the MUFOLD-MD group (TS404_5). By visual assessment this model was ranked as best model by all three assessors. It predicts the α-helices very well, with their orientation and position resembling the target quite well. For the β-sheet strands, the prediction is not as good, that is, it predicts three strands instead of six, and, in fact, strands 5 and 6 were predicted to be α-helices. It is therefore rated as a fair, rather than excellent, best model. The next highest scoring model is from the Handl-Lovell group (TS029_3). In the visual assessment, this model was chosen by the three assessors to be in the top three models. Because of some imperfections in the helix orientations this model is considered a less good model in comparison to TS404_5 (see Fig. 7). In addition, there is a significant drop in GDT_TS score from 31.85 to 25.12 between the top two models.

T0416-D2 (FM; S); PDB 3d3q

This short domain (57 residues) is a bundle of four differently sized helices in an up-and-down topology. Many groups built quite good models with only minor imperfections, such as orientation or tilting of one of the helices (see Fig. 4). A few models stood out by visual inspection, as well as by GDT_TS score and OK_Rank: TS404_2 from group MUFOLD-MD with two top-model votes and thus considered best model, and a near-identical trio with three votes in the top 3 from McGuffin (TS379_2), Zhang-Server (TS426_5) and MULTICOM (TS453_1), for which TS426_5 was considered the originating server. Three additional models also predicted this target well, that is, TS425_5, TS166_5, and TS340_5, but each received only one vote in the visual inspection.

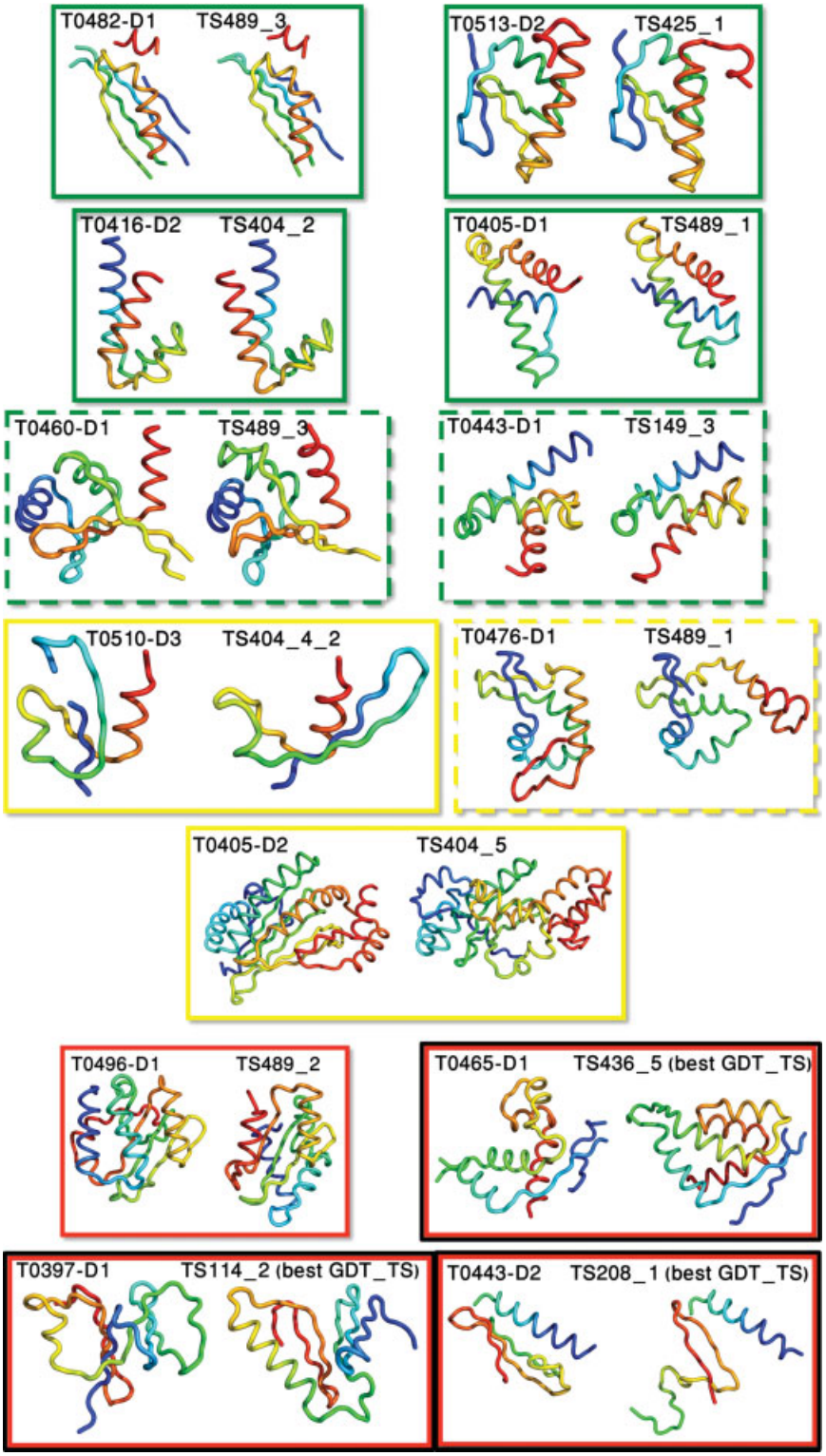

Figure 4

All FM and FM/TBM targets with their corresponding best models. Targets and best models are arranged according to the best model quality: excellent models (framed in green), fair models (framed in yellow), and poor models (framed in red, see text for more details on model classification of excellent, fair, and poor). The assessors could not choose even a single good model for T0465-D1, T0397-D1, and T0443-D2 hence the best GDT_TS scoring models are shown (framed in black). FM/TBM targets are displayed with a dotted frame.

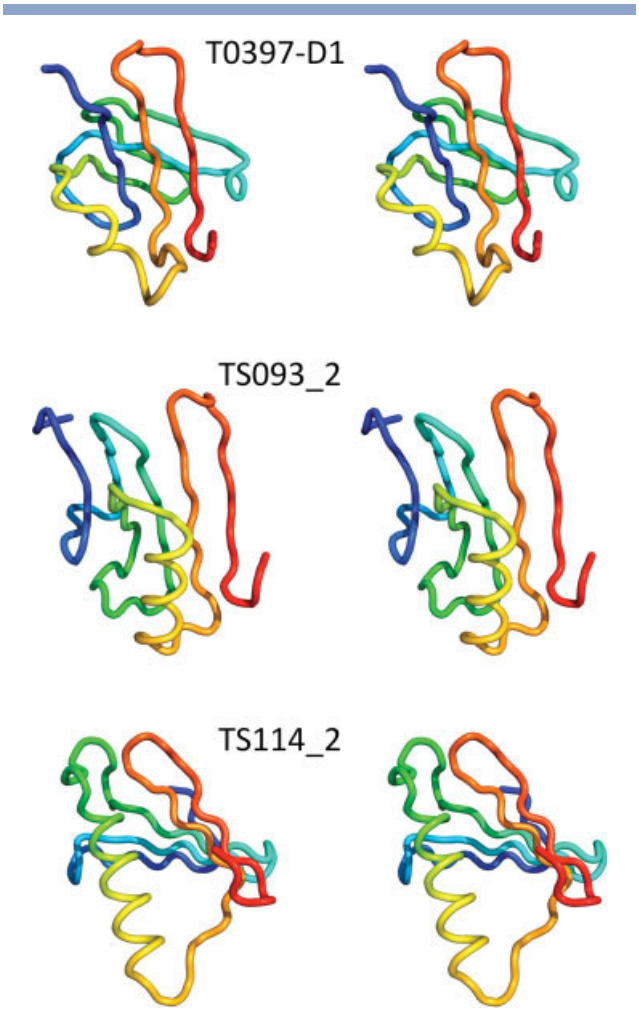

Figure 5

Structure of T0397-D1, which is classified as a new fold, was very difficult to predict due to its unusual topology. Although the model TS114_2 showed poor correlation to experimental structure it received the highest GDT_TS score (35.97). On the other hand, the model TS093_2 showed relatively better agreement with the experimental structure, it was only 38th in the GDT_TS list (score 30.79).

T0443-D1 (FM/TBM; H/S); PDB 3dee

This is an all-α domain with three main helices and two short helices connecting them. Two models stood out, both from A-TASSER (TS149_3, TS149_5). These models are quite similar to each other, although minor differences made TS149_3 the top-ranking model for GDT_TS score, OK_Rank, and visual inspection by all three assessors. The second model (TS149_5) was consistently chosen in the top three. Many groups did well in predicting the two first main helices but missed the third one, probably due to contact with another helix from the second domain.

T0443-D2 (FM; H/S); PDB 3dee

This domain is an α + β structure, with one long α-helix followed by three antiparallel β-strands. Many groups were able to predict the long helix and the last two β-strands. However, these groups mistakenly predicted the first β-strand to be an α-helix. None of the models had a good orientation and accurate position of the secondary structural elements. Therefore, independently, all three assessors felt that there was no good model for this target. Model TS208_1 has the highest GDT-TS, MAMMOTH-Z, and Dali-Z scores and is thus considered the best available model, but it is quite non-compact and thus of poor quality (see Fig. 4).

T0460-D1 (FM/TBM; H/S); PDB 2k4n

This NMR determined domain consists of a four-stranded β-sheet and three α-helices. Residues 50–71 were trimmed from the target, since they form a disordered loop. Many groups did fairly well in predicting the first two helices (the part before the disordered loop), yet missed the right orientation of the second part, the two β-strands and helix near the C-terminus. The top-scoring model in all measures (GDT_TS score, OK_Rank, and ranked as best model by all three assessors) is from the Baker group (TS489_3). This is an excellent model with the three helices at nearly the correct position and orientation, with minor imperfections in the last helix, which is a bit bent relative to the target structure. The second high scoring model is from the Jones-UCL group (TS387_1). This model also has high GDT_TS score and OK_Rank (Table I), and in the visual assessment it was chosen by all three assessors to be in the top two models. It is quite similar to the top-scoring model except in some connecting loops.

T0465-D1 (FM; H/S); PDB 3dfd

This domain consists of five α-helices of different sizes and two β-strands. The 10 models with the highest GDT_TS scores are virtually identical: Pcons_dot_net (TS436_5), BAKER-ROBETTA (TS425_5), MULTICOM (TS453_4), Zico (TS299_3), ZicoFullSTP (TS196_4), ZicoFullSTPFullData (TS138_3), and MUFOLD (TS310_2). The originating free model for this cluster presumably came either from server 436 or server 425. These models all resemble the target structure in predicting the secondary structural elements; however, there is a shift, of about the width of one α-helix, of the helices relative to the target. In addition, some of the helices are misoriented (see Fig. 4). In the OK_Rank, this cluster obtained poor ranks (between 8 and 22); however, although the visual assessment rated this cluster in the top two, that is, in agreement with the GDT_TS score, the assessors felt that this cluster yielded a relatively poor model.

T0476-D1 (FM/TBM; H/S); PDB 2k5c

This NMR structure consists of a helix bundle topped with two β hairpins, which form a metal binding site

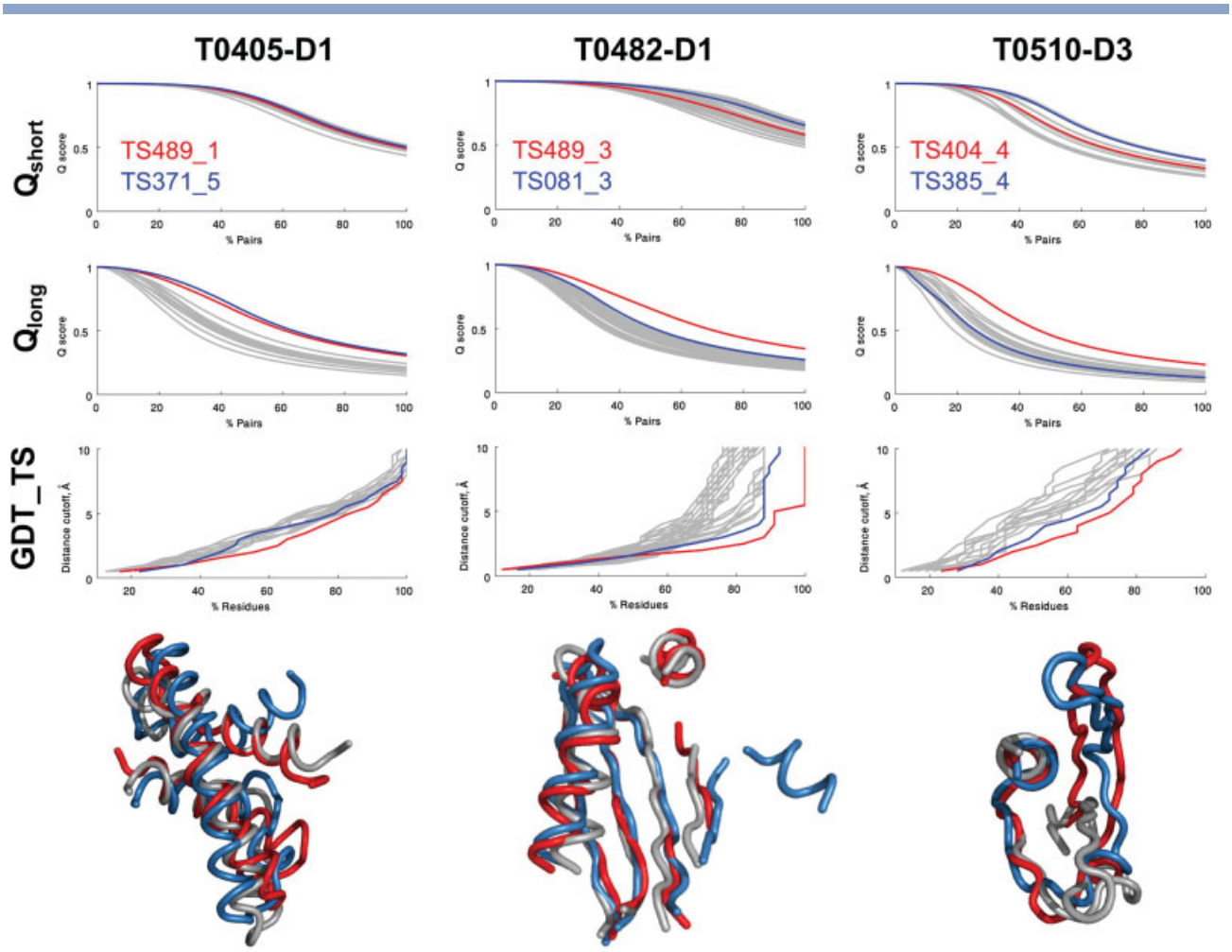

Figure 6

Q_{short}, Q_{long}, and GDT_TS plots for targets T0405-D1, T0482-D1, and T0510-D1. The gray lines correspond to models ranked at the top 15 by the OK_Rank. The red and blue lines correspond to the best models chosen by the visual inspections. The partial agreement that is often found between Q_{short}, Q_{long}, and GDT_TS reflects the complexity of the assessment and the need for more than a single measure as well as a visual examination of the best structures. The structures of the two best models (red and blue) are compared to the target (grey).

(zinc in the target structure). This target had a template (2q5h_A) covering the first 60 residues of the structure that is fairly conserved, however only two groups reported using it as a template.[9] Models by these two groups, DBAKER (TS489_1) and MUFOLD-MD (TS404_2), obtained the top GDT_TS and OK_Rank. Model TS489_1 (Fig. 4) was also selected as the top two by visual assessment, whereas only two assessors selected model TS404_2. Both models accurately predict the position of the helices, whereas model TS404_2 suggests a bit more accurate orientation of these helices. On the other hand, model TS489_1 was more accurate in the length and orientation of the two hairpins. In addition, both models inaccurately position an additional short helix between the second hairpin and the last helix. The last 15 residues were the most difficult to predict, and both models failed to do so with errors in the position, orientation, and secondary structure prediction. Although this part in the target structure is a coil, model TS489_1 predicted it as an α-helix, whereas in model TS404_2 it was predicted as β-strands.

T0482-D1 (FM; H/S); PDB 2k4v

This NMR structure consists of four antiparallel β-sheet strands, together with a short and a long helix that in the target were connected by a disordered loop trimmed in the domain definition process. The best model is from the Baker group (TS489_3) with the highest scores on all measures (GDT_TS, Q_{long} plots, RMSD, and assessor votes), clearly reflecting its excellent quality (see Figs. 6 and 8). This model was also the only one to predict all structural features in the right position and orientation. The model assessed as the second best is

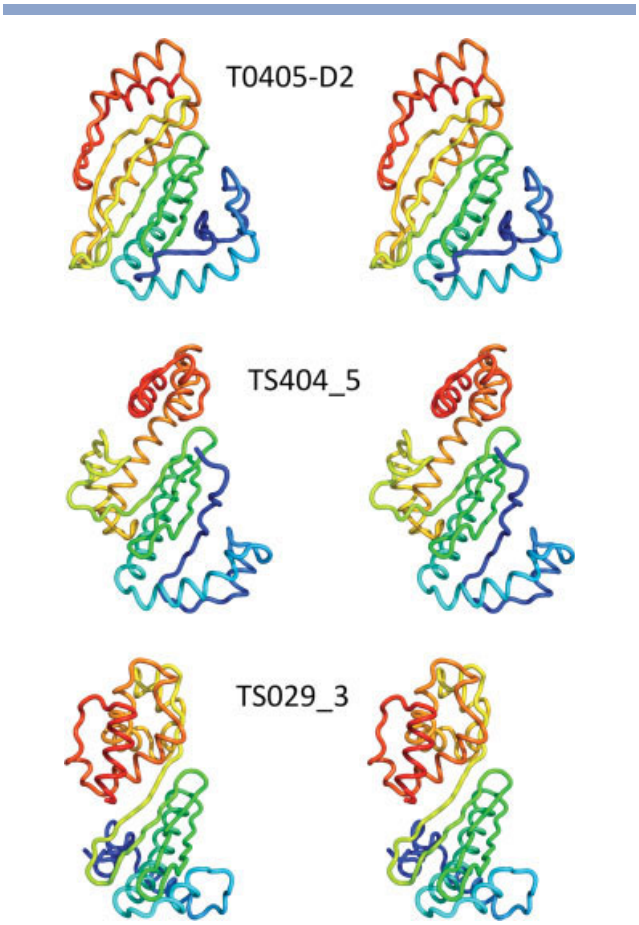

Figure 7

Models that have different secondary structures for the same part in the target. For the last part of the domain, both of the *best* models were incorrect. Model TS404_5 predicted this region as a helix and model TS029_3 as β-strand, whereas in the target this part has no secondary structure (coil).

from the Chicken George group (TS081_3). This model also obtained high scores and was the second in all measures, except RMSD. However, it failed in positioning and orientation of the short helix (polarity, i.e., it had the N⇗C pointing in the wrong direction). The secondary-structure predictions of many groups were quite correct (see the Q_{short} plots, Fig. 6), many groups did fairly well in the positioning and orientation of the β-strands, and some could also predict the long helix with minor imperfections. Although some groups reported using templates for the prediction, these template-based models were quite poor.

T0496-D1 (FM; H/S); PDB 3do9

This domain consists of five differently sized helices and four antiparallel β-strands. It was a difficult target, as indicated by the relatively low GDT_TS scores and the results of the visual assessment, where there was no agreement between the suggested models and moreover, one assessor suggested that none of the models were good. Models 1 and 2 on GDT_TS score were from the Baker group (TS489_2 and TS489_3), which also obtained one vote in the visual assessment; TS489_2 also had the highest Dali-Z score. Another model that was voted for is from the Poing group (TS186_4), which ranked second in the OK_Rank (TS489_2 was third). Both groups correctly predicted some of the secondary structural elements, yet there were errors in the sheet topology, and inaccuracies in the positions and orientations of the helices, which made it very difficult to visually inspect (see Fig. 4).

T0510-D3 (FM; S)

This short domain (43 residues) includes two antiparallel β-strands and one α-helix connected by a long loop. A number of groups predicted the secondary structural features well; however, they failed to place and orient the elements correctly. Some groups did well in the prediction of the first part of the structure (β-strands), whereas others did well in the last part only (helix). The best model by many measures is from the MUFOLD-MD server (TS404_4_2) (see Fig. 9). This model was in the top two of the visual assessment and with the highest ranks of GDT_TS and the OK_Rank. It stands out by both the Q_{long} and the GDT_TS measures (but there are models with better Q_{short}) (Fig. 6). Albeit the first part of the domain is rather misoriented relative to the rest, this group had a fairly good prediction for the last part including the majority of the connecting loop. ABIpro (TS340_3) and PSI (TS385_4) models align perfectly with each other with high scores and were rated the second-best models. On the other hand, RAPTOR (TS438_1) and another model by MUFOLD-MD (TS404_1_2) oriented and placed well the β-strands well, but failed to place the helix (see Fig. 9).

T0513-D2 (FM; S); PDB 3doa

This domain contains four antiparallel β-strands and two α-helices (Fig. 10). The top GDT_TS and OK_Rank models (∼28) are all virtually identical and are treated as one cluster (see Fig. 11). It includes models from two servers that could have acted as the original template for the others; the TS425_1 model from BAKER-ROBETTA was submitted 10 h earlier than the five models from GS-KudlatyPred (Andriy Krystafovych, personal communication) and is therefore judged to be the original free model. The models in this cluster are excellent predictions, with just minor imperfection in the last helix (residues 62–82) (Fig. 4). Other groups succeeded in getting the correct position of this helix (e.g., FEIG TS166_4 and SAMUDRALA (TS034_3), however, they unfortunately failed in predicting other features of the target structure.

Table I
The Ranking Based on GDT_TS, OK_Rank, and Assessor Votes, for the Models Inspected Visually

		H/S		
Target	Model	# Top 3 selections	GDT_TS rank	OK_Rank
T0397-D1	TS114_2	1	1	1
	TS479_2	1	2	2
	TS138_3	1	3	3
	TS453_2	1	4	4
	TS299_4	1	4	8
	TS196_3	1	4	8
	TS178_5	1	4	10
	TS178_4	1	4	10
	TS178_3	1	4	17
	TS138_5	1	4	7
	TS453_1	1	6	11
	TS182_1	1	7	9
	TS093_2	1	15	29
T0405-D1	TS489_1	3	1	1
	TS371_5	3	4	3
	TS387_5	1	2	2
T0405-D2	TS404_5	3	1	1
	TS29_3	3	2	18
	TS114_1	2	4	23
	TS46_3	2	8	13
	TS479_1	1	11	6
	TS442_1	1	11	6
	TS479_4	1	7	7
	TS310_1	1	13	45
	TS29_5	1	4	29
T0443-D1	TS149_5	3	3	6
	TS149_3	3	1	1
	TS404_5	1	2	13
	TS114_3	2	4	4
	TS119_1	1	5	2
	TS453_3	1	6	6
	TS425_1	1	6	8
	TS325_2	1	6	10
	TS310_2	1	6	8
	TS299_1	1	6	8
	TS196_1	1	6	8
	TS138_2	1	6	8
	TS46_4	1	3	7
	TS46_2	1	6	9
	TS46_1	1	6	5
T0460-D1	TS489_3	3	1	1
	TS387_1	3	2	3
T0465-D1	TS436_5	3	1	22
	TS425_5	3	1	19
	TS453_4	3	2	17
	TS299_3	3	2	8
	TS299_2	3	2	11
	TS196_4	3	2	9
	TS196_3	3	2	11
	TS138_3	3	2	13
	TS138_2	2	2	13
	TS310_2	3	3	20
	TS71_1	1	3	2
	TS207_4	1	3	15
	TS434_1	1	6	1
T0476-D1	TS489_1	3	1	2
	TS404_2	2	2	1
	TS70_1	1	8	8
T0482-D1	TS489_3	3	1	1
	TS81_3	3	2	2
T0496-D1	TS186_4	1	9	2
	TS207_5	1	16	3

Table I
(Continued)

		H/S		
Target	Model	# Top 3 selections	GDT_TS rank	OK_Rank
	TS489_2	1	1	11
	TS489_3	1	2	7
	TS387_1	1	3	50
T0513-D2	TS453_1	3	Identical	
	TS453_2	3		
	TS453_3	3		
	TS453_4	3		
	TS279_1	3		
	TS279_2	3		
	TS279_3	3		
	TS279_4	3		
	TS279_5	3		
	TS299_2	3		
	TS299_4	3		
	TS379_1	3		
	TS379_3	3		
	TS379_4	3		
	TS196_2	3		
	TS196_3	3		
	TS196_5	3		
	TS138_2	3		
	TS138_4	3		
	TS138_5	3		
	TS425_1	3		
	TS340_1	3		
	TS340_2	3		
	TS340_3	3		
	TS340_4	3		
	TS340_5	3		
	TS124_3	3		
	TS404_2	3		
T0510-D3	TS404_4_2	3	1	1
	TS340_3	3	2	2
	TS385_4	2	2	2
	TS438_1	1	4	24
	TS404_1_2	1	4	6
	TS340_2	1	4	6
	TS404_4_2	1	1	1
	TS404_3_2	1	3	8
T0416-D2	TS404_2	2	1	1
	TS379_2	3	3	2
	TS426_5	3	3	2
	TS453_1	3	4	4

The number of top selections indicates how many assessors ranked the model as a top model based on visual inspection.

α-helical versus β-sheet predictions

To evaluate the quality of predictions of α helical and β-sheet regions in the models, we used versions of the Q score that incorporate only inter-helical or inter-strand pairwise distances (helical and strand stretches were assigned using the DSSP program[21] for classifying secondary structure), and are respectively called $Q_{\alpha\text{-helix}}$ and $Q_{\beta\text{-sheet}}$. These calculations were implemented for each model that was ranked at the top 15 by the GDT_TS score. The plots of $Q_{\alpha\text{-helix}}$ and $Q_{\beta\text{-sheet}}$ as a function of fraction of pairwise distances depict the mean of the cor-

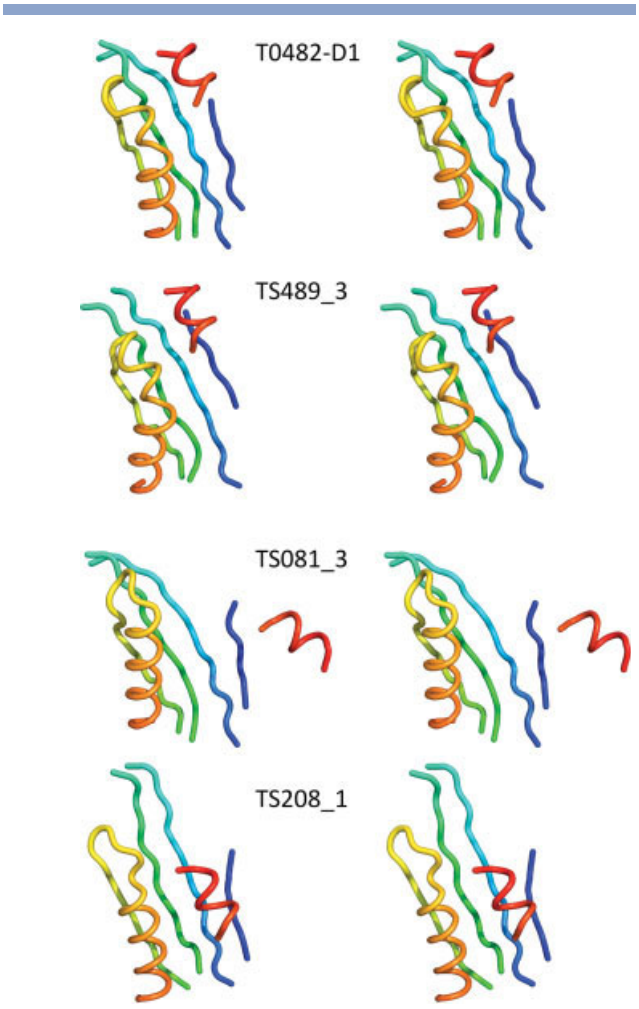

Figure 8

A good model with low GDT_TS and OK_Rank for Target T0482-D1. Model TS489_3 is clearly the best model by all measures. The second best model chosen was TS081_3; however, TS208_1 is arguably as good, since it positioned the small helix correctly. The assessors were not aware of TS208_1, since it had low GDT_TS and OK_Ranking.

responding Q of the top 15 models as well as the standard deviation (Fig. 10). Surprisingly, for most targets the registration of β-strands was better predicted than the packing of the α-helices. This results presumably from the fact that within a given sheet the inter-strand distances are controlled by hydrogen bonding, and only between separate sheets is the packing more variable. For targets T0482-D1 and T0513-D2 (both have excellent models, see Fig. 10), it was found that $Q_{\beta\text{-sheet}}$ is quite high even when all the pairwise distances involve in the inter-strand interactions are taken into account. This illustrates that the β-sheets are very well predicted. In contrast, the accuracy of predicting the helix packing is more limited even when the helices themselves (e.g., their length and position in the sequence) are correct. In T0482-D1, the two helices were very poorly predicted and in T0513-D2 they were reasonably well predicted yet the orientation of the two helices was shifted. In target T0496-D1 (has only poor models; see Fig. 10), both helices and sheet are poorly predicted, yet $Q_{\beta\text{-sheet}}$ is higher than $Q_{\alpha\text{-helix}}$, indicating better predictions for inter-strands over the inter-helices interactions (Fig. 10). The higher $Q_{\beta\text{-sheet}}$ scores could have the advantage of offsetting the somewhat unfair advantage of helices in most other scores (such as GDT-TS) just because they include more residues.

Cluster of very similar models

An important issue that was raised during the assessment of the predictions of the FM and the FM/TBM targets is the existence of a cluster of extremely similar superimposable models from multiple groups, which show near-exact coordinate matches for Cα atoms distant in sequence and structure. The targets T0397-D1, T0416-D2, T0443-D1, T0465-D1, and T0513-D2 include clusters of 10, 10, 8, 10, and 26 models (Fig. 11). Running these targets on Dali reveals that there are no templates (which were missed during the target assignment by the CASP organizers) that might be used in the prediction of these targets. It is therefore likely that different

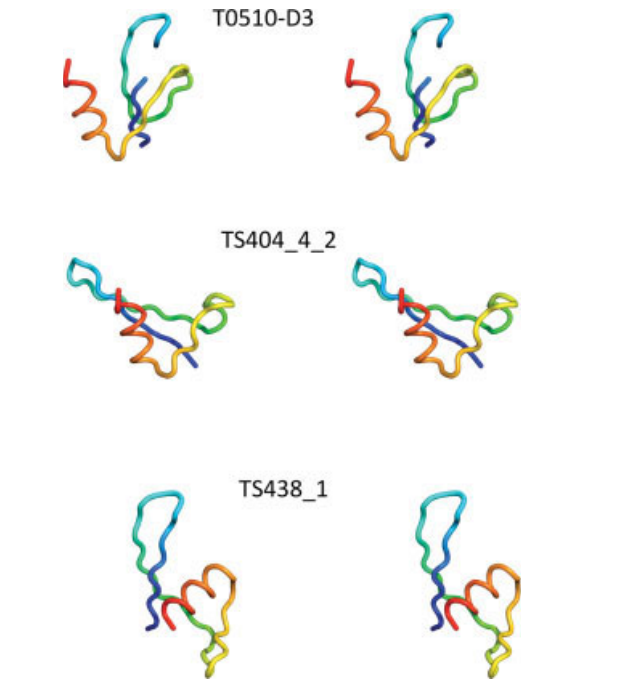

Figure 9

Successful predictions for parts of a domain. Although target T0510-D3 is quite short, none of the models were able to provide a good prediction for the whole domain. Some models did well in the prediction of the first part (e.g., TS438_1), whereas others succeeded in predicting the second part (e.g., TS404_4_2).

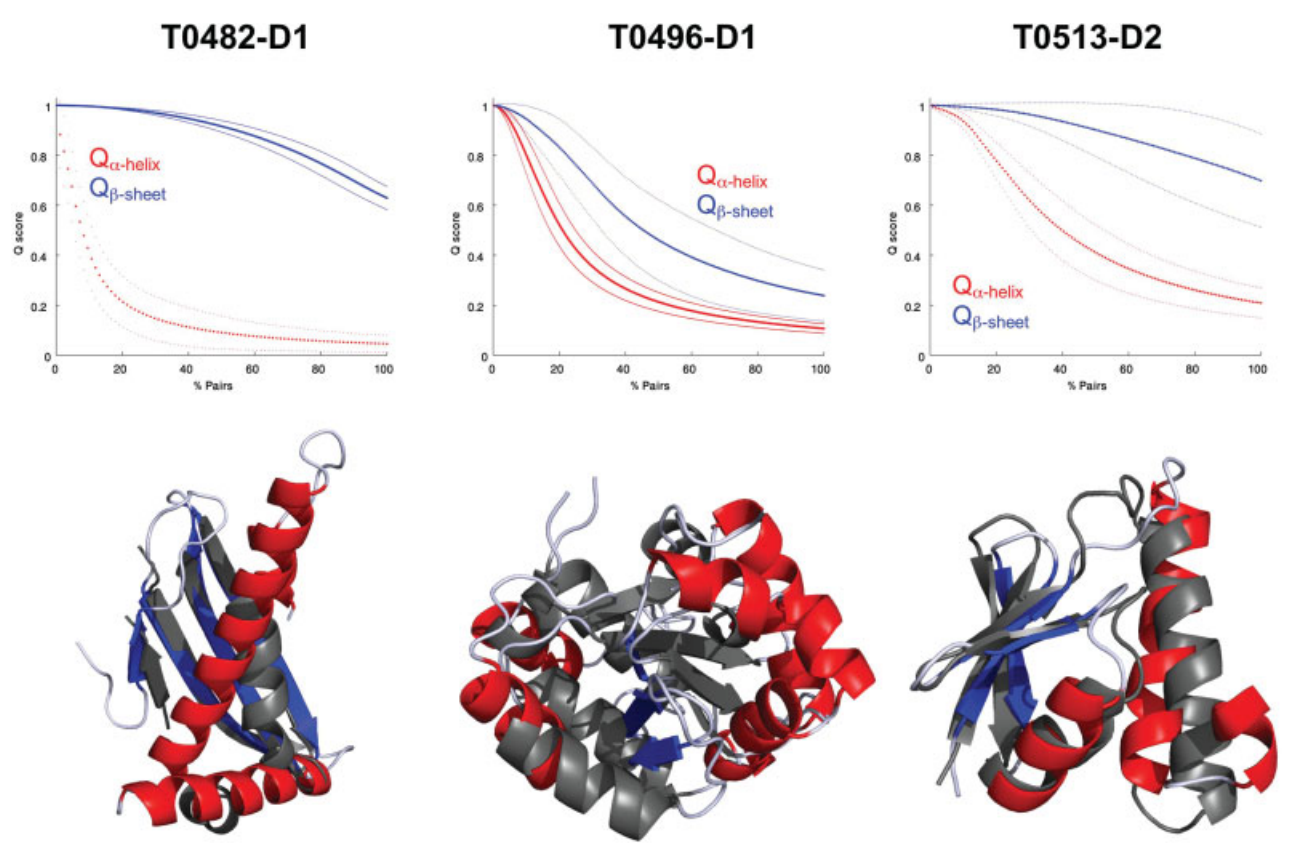

Figure 10

$Q_{\alpha\text{-helix}}$ and $Q_{\beta\text{-sheet}}$ for targets T0482-D1, T0496-D1, and T0513-D2. The Q measures were calculated for the models that were ranked at the top 15 by the GDT_TS score, resulting in about 40 models for each target. The large and small dots correspond to the mean value of Q and the standard deviation for the selected models. For illustration, one of the models of each target is shown together with the target (in grey). α-helices and β-strands are shown by red and blue, respectively.

groups used the same model (or models) released from prediction servers. Accordingly, each of the clusters of these five targets includes at least one server model which could have done the original FM prediction that then after its public release acted as a template for the other groups. There is absolutely nothing wrong with this prediction approach. Actually, it is a very valuable achievement to recognize good starting models. Yet, this is not a template-free modeling. Therefore the existence of cluster of similar structures suggests that each group submitted such a model that was predicted using a server may not be treated as it was independent. However, downscoring groups that used models released from prediction servers (and crediting the server) is complicated and also required the identification of that server. Although we think that it is important to take into account in the scoring scheme the existence of near-identical models, in the assessment of CASP8 targets we have not implemented such an approach.

Best performing groups

We have ranked the performance of the different groups after the integration of all three assessors' votes (Supporting Inforamation Table 1). As described in the methods section, scoring scheme M provides data about the number of different targets each group has successfully modeled. Table II shows that MULTICOM is ranked first with votes for seven out of the 13 targets they submitted, the MUFOLD-MD server scored 6/13, and in the third rank DBAKER scored 5/10, BAKER-ROBETTA server scored 5/13, and Zico and ZicoFullSTPFullData scored 5/13 as well. Another group worth mentioning with a high success rate was the Keasar group with 4/10. These data show that the highest percent of high quality models per target attempted is 54% from the MULTICOM group. It should be noted, however, that MULTICOM, Zico, and several other groups start from the released server models; therefore they are not doing template-free modeling in the strict sense, but have proven

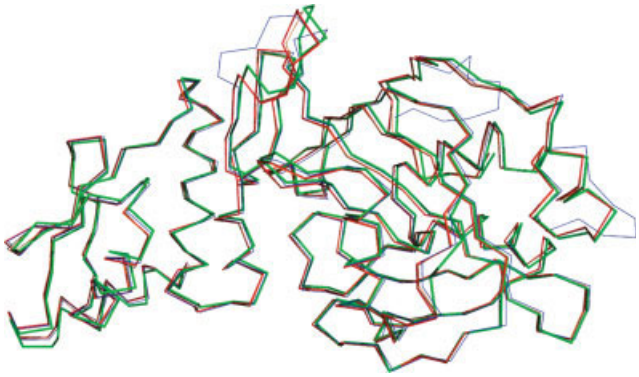

Figure 11

Clustering in T0513-D2. For this target, a cluster of 26 nearly identical models, from eight different groups could be identified. Only one structure from each of these eight groups is shown. Line thickness is proportional to the number of groups that submitted identical models (green for groups 138, 196, 299; red for groups 379, 425; black for group 279; blue for group 340; orange for group 453).

highly successful at identifying good server models to act as further templates for this category of target.

Another important factor in ranking the groups is the total number of "high-quality" models, captured by Scoring Scheme A (Table III). Using this scheme, MULTICOM and ABIpro are ranked as number 1 with 24 votes, ZICO and ZicoFullSTPFullData scored 20 and the two servers MUFOLD-MD and GS-KudlatyPred were ranked third with 18 votes.

Finally, Table IV shows the number of best models per group. DBAKER dominates this category with five best models: excellent ones for T0405-D1, T0460-D1, and T0482-D1; a fair one for T0476-D1, and a poor one for T0496-D1. The MUFOLD-MD server has three best models: an excellent one for T0416-D2, and fair ones for T0405-D2 and T0510-D3. A-TASSER has an excellent best model for T0443-D2. The BAKER-ROBETTA server produced an excellent model for T0513-D2 and shares with Pcons_dot_net the probable responsibility for a poor best model on T0465-D1. The Wolynes group produced a poor best model for T0397-D1, and MidWay-Folding a poor best model for T0443-D2.

DISCUSSION

There is no doubt that in recent years many obstacles have been removed on the long and elusive way toward deciphering the protein-folding problem.[16,22] The current understanding of the physics of protein folding[22–25] is quite advanced, and this is nicely reflected by numerous collaborative researches of experimentalists and theoreticians aiming at providing an integrated atomistic view of folding mechanisms.[26,27] There have even been commentaries written that the protein folding research field is on the verge of tackling the complete problem.[28] In the case of free model prediction, as evaluated in this CASP, impressive successes have been achieved, yet the problem is far from being solved.

From the visual assessment of 10 FM and 3 FM/TBM targets, from all the groups that participated in CASP8, only six targets had excellent models, of which two were FM/TBM. Three targets were judged to be fair and four as poor (Fig. 4). It should be noted that, in fact, for most of the targets with an excellent model, only a small subset of the groups submitted models which were indeed excellent, and most models were rather far from predicting the 3D structures of the targets. Moreover, Table II clearly shows that no successful group had more than ~50% of the targets ranked in the "top 3" by at least one of the assessors.

Table II

Groups Performance: Scoring Scheme M

Group name	Group index	Scheme_M score	Number of submitted targets
MULTICOM	453	7	13
MUFOLD-MD (s)[a]	404	6	13
DBAKER	489	5	10
BAKER-ROBETTA (s)	425	5	13
Zico	299	5	13
ZicoFullSTPFullData	138	5	13
ZicoFullSTP	196	4	13
Keasar	114	4	10
Jones-UCL	387	3	13
ABIpro	340	3	13
MUFOLD	310	3	9
RBO-Proteus	479	2	13
Pcons_dot_net (s)	436	2	13
fams-ace2	434	2	13
McGuffin	379	2	13
GeneSilico	371	2	10
POEM	207	2	10
Bates_BMM	178	2	10
Zhang	71	2	10
SAM-T08-human	46	2	10
LevittGroup	442	1	10
RAPTOR (s)	438	1	13
Zhang-Server (s)	426	1	13
PSI (s)	385	1	13
FALCON (s)	351	1	13
Bilab-UT	325	1	10
GS-KudlatyPred (s)	279	1	13
Poing (s)	186	1	13
METATASSER (s)	182	1	13
FEIG (s)	166	1	13
A-TASSER	149	1	13
POEMQA	124	1	13
SAINT1	119	1	9
Wolynes	93	1	6
Chicken_George	81	1	10
Fleil	70	1	10
TASSER	57	1	13
Handl-Lovell	29	1	8

[a](s) indicates Server. Please see Tables III and IV.

Table III
Groups Performance: Scoring Scheme A[a]

Group name	Group index	Number of top 3 votes	Number of submitted targets
MULTICOM	453	24	13
ABIpro	340	24	13
ZicoFullSTP	196	20	13
ZicoFullSTPFullData	138	20	13
MUFOLD-MD (s)[b]	404	18	13
GS-KudlatyPred (s)	279	18	13
Zico	299	16	13
DBAKER	489	14	10
McGuffin	379	12	13
BAKER-ROBETTA (s)	425	10	13
SAM-T08-human	46	7	10
MUFOLD	310	6	9
A-TASSER	149	6	13
Keasar	114	6	10
Jones-UCL	387	5	13
Bates_BMM	178	5	10
RBO-Proteus	479	4	13
Pcons_dot_net (s)	436	4	13
GeneSilico	371	4	10
Handl-Lovell	29	4	8
fams-ace2	434	3	13
Zhang-Server (s)	426	3	13
PSI (s)	385	3	13
POEMQA	124	3	13
Chicken_George	81	3	10
FALCON (s)	351	2	13
Bilab-UT	325	2	10
POEM	207	2	10
Zhang	71	2	10
Fleil	70	2	10
TASSER	57	2	13
LevittGroup	442	1	10
RAPTOR (s)	438	1	13
METATASSER (s)	182	1	13
FEIG (s)	166	1	13
SAINT1	119	1	9
Wolynes	93	1	6

[a]This table was constructed for all targets except T0513-D2 that had many identical models.
[b](s) indicates Server.

In the visual assessment process independently performed by three assessors, no assessor could choose even one good model for target T0443-D2. For targets T0397-D1 and T0496-D1 one assessor (a different assessor for each target) could not choose a good model. In these targets and a few other "difficult" targets the visual assessment was extremely difficult and problematic due to the low resemblance between target and models and we felt that the task of visual assessment became more qualitative and subjective.

It is important to note that the correlation coefficient between all four scores comprising the OK_Rank is around 0.8, indicating that although there is good correspondence between all scores, each score emphasizes different properties of the models (e.g., secondary vs. tertiary structure) and thus provides a balanced way to narrow down the model list that was assessed visually. GDT_TS and Q_{long}, which are highly correlated (the average correlation coefficient for the 13 targets is 0.80 $\pm$ 0.13), are useful tools in narrowing down the model list for each target. In addition, the GDT_TS and OK_Rank correlate well with our visual assessment (Table I) (the averaged ranking based on GDT_TS and OK_Rank of the top three models selected by visual inspections is 4.0 $\pm$ 3.2 and 8.9 $\pm$ 9.2, respectively), despite some problems that have been discussed in previous CASP experiments[29] and current problems with T0397-D1. We emphasize that the advantage of the Q score is that it overcomes the need for structural alignment and it can be easily manipulated and therefore can be used to compare separately various parts of the protein structure. We found that correct interactions between residues with a large separation in sequence (i.e., high Q_{long}) are crucial for good predictions. Often, prediction with high Q_{long} had relatively low Q_{short}, suggesting that calibration the weighting of local and distant pairwise interactions may improve structure predictions. In particular, we found that inter-strand interactions are better predicted than inter-helical ones, highlighting the need to improve prediction of helix packing.

Ranking the groups is far from trivial, since each group could submit up to five models per target, might not submit models for all targets, and to complicate things even further, for a few targets many models from the same group were good whereas for most targets only one model was good. These factors made us employ two scoring schemes, each emphasizing different features as an aid in pinpointing the best performing groups. As Scoring Scheme M highlights the number of targets for which a group had high quality models and Scoring Scheme A highlights the total number of high quality models per group, one can compare the two and notice that MULTICOM is ranked first by both schemes; MULTICOM had a number of good models (3.4 on average) for seven out of the 13 targets. This clearly shows the merits of this group, but Scheme M does not provide

Table IV
Number of Best Models by Group

Group	Number of BEST models
DBAKER	5
MUFOLD-MD (s)[a]	3
BAKER-ROBETTA (s)	1
Keasar	1
A-TASSER	1
MidWayFolding	1
GS-KudlatyPred (s)	1
MULTICOM	1
Pcons_dot_net (s)	1
Zico	1
ZicoFullSTP	1
ZicoFullSTPFullData	1

[a](s) indicates Server.

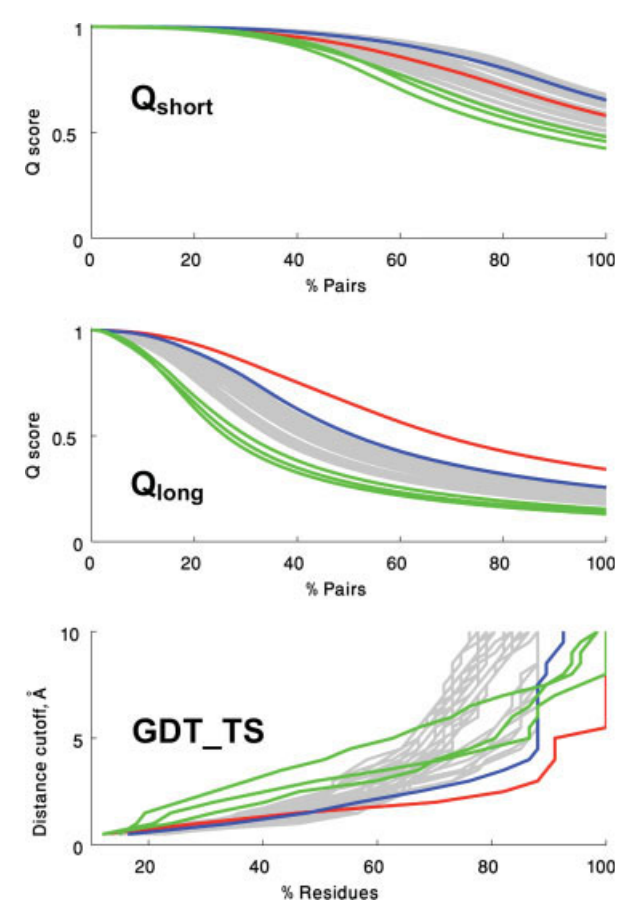

Figure 12

Q scores and GDT plots for T0482-D1. The grey lines correspond to models ranked in the top 15 by the OK_Rank. The red and blue lines correspond to the best models chosen by visual inspection (TS489_3 and TS081_3, respectively). The green lines correspond to models TS208_1, TS387_2, and TS020_1 that are worse than the two best models for most of their length (see Q_{short} and Q_{long}, upper and medium panels, and cumulative GDT-TS plot) but in ~10% of its length (probably corresponding to the short helix discussed in the text) they are superior. For example, for model TS208_1, 100% of the residues have deviation smaller than 8 Å while for model TS081-3 about 90% of the residues have similar deviation.

insight about best models. As described in the results section, in many cases there was a considerable difference in the quality of the best model and the second best. Ranking the groups by the best model highlights another group: the DBAKER group had five best models and MUFOLD-MD had three, whereas MULTICOM had only one best model. We note that the existence of large cluster of near-identical models suggests that a more elaborate scoring scheme might be applied to downweight because of the usage of a "template" which is provided most likely by a server. Downweighting (or sharing scores among models) and crediting the server that started the cluster are important in the assessment of FM targets to evaluate progress in structure prediction and should be considered in future CASP experiments.

As mentioned earlier, only models with top scoring GDT_TS and OK_Rank were assessed. The GDT_TS and OK_Rank are overall scores for the fit of the entire model to the target. It is interesting to note that a few models for T0482-D1 had GDT_TS and OK_Rank scores below the cutoff, but at first glance, in fact, do not look too bad by visual inspection, especially in the short helix at the C-TERM of this structure. This was brought to our attention for model TS208_1 (MidwayFolding) that we had not assessed due to its low ranking of both of these scores. Once it was brought to our attention, we reinspected it. It is clear that this model, as well as several others, had relatively poor overall GDT_TS scores due to local mispositioning of a large portion of the structure relative to the target. However, the short C-terminal helix looked better than in many models that indeed passed the cutoff of the GDT_TS and OK_Rank. This success in a limited part of the prediction may be clearly visualized by examination of the GDT-plots (Fig. 12) in which these models are better for the far right portion of the cumulative GDT-plot, corresponding to fitting the most residues in the 10 Å cutoff; however, these models are by no means the best for considerable portions of the structure. Thus these cumulative GDT-plots are very useful to aid in visual inspection for FM models, which often are difficult to quantitatively rank by other means, and in pointing out fragments of models that are markedly better.

Finally, the CASP8 FM and FM/TBM experiment included only 13 targets. This small number of targets, with 11 in the size range of 44–87 amino acids, makes it difficult to obtain statistically significant conclusions on the current CASP FM experiment. Moreover, if one wishes to compare the performance of the free modeling

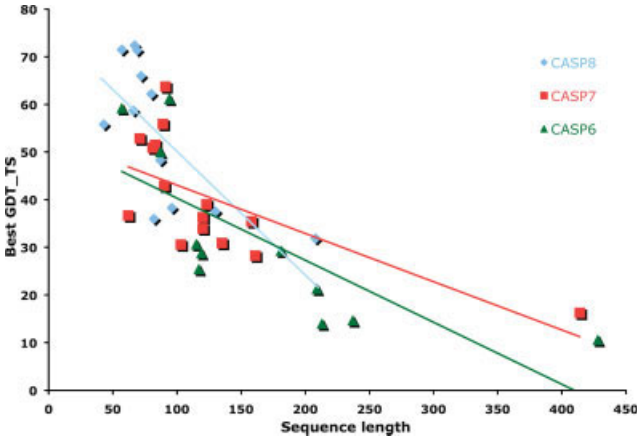

Figure 13

Maximal GDT_TS scores for FM targets in CASP 6–8 as a function of target lengths.

predictors throughout different CASP experiments the same problem is relevant and also the selection rules for the FM domains have changed, so that these comparisons are very problematic as discussed by Noivirt *et al.*[30]

Thus for template-free model prediction, as evaluated in this CASP, successes have been achieved for most targets, and it appears that the best models' GDT_TS scores have improved in comparison to CASP 7 and 6 (Fig. 13). However, a great deal of research is still required in both improving the existing methods and in development of new approaches. In addition to better sampling of the fold space, which may be feasible thanks to improvement in computer capabilities and particularly by the use of graphics processing units (GPU), advancing the underlying physical principles of structure prediction schemes will offer an important venue for improvements. Incorporating fundamental physical concepts of folding mechanism achieved in the last two decades (in particular, the funnel-shaped energy landscape) may advance quality and convergence of predictions as well as reduce the need for exhaustive sampling for the native state.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Jane S. Richardson for her careful reading of this manuscript and her very constructive suggestions, and the members of the Protein Structure Prediction Center for all their help in preparation of this manuscript. J.L.S. is the Morton and Gladys Pickman Professor of Structural Biology and Y.L. is the incumbent of the Lilian and George Lyttle Career Development Chair.

REFERENCES

1. Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction - round VII. Proteins 2007;69:3–9.
2. Moult J, Comparative modeling in structural genomics. In: Sussman JL, Silman I, Eds. Structural proteomics and its impact on the life sciences. Singapore: World Scientific Publishing Company, 2008;121–134.
3. Jones TA, Thirup S. Using known substructures in protein model building and crystallography. EMBO J 1986;5:819–822.
4. Jones DT. Successful *ab initio* prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. Proteins Suppl 1997;1:185–191.
5. Unger R, Harel D, Wherland S, Sussman JL. A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins 1989;5:355–373.
6. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 1998;281:565–577.
7. Jauch R, Yeo HC, Kolatkar PR, Clarke ND. Assessment of CASP7 structure predictions for template free targets. Proteins 2007;69:57–67.
8. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucl Acids Res 2003;31:3370–3374.
9. Tress ML, Ezkurdia I, Richardson JS. Target domain definition and classification in CASP8. Proteins 2009;77(Suppl 9):10–17.
10. Sussman JL, Silman I, editors. Structural proteomics and its impact on the life sciences. Singapore: World Scientific Publishing Company, 2008.
11. Chandonia J-M, Brenner SE. The impact of structural genomics: expectations and outcomes. Science 2006;311:347–351.
12. Shi S, Pei J, Sadreyev RI, Kinch LN, Majumdar I, Tong J, Cheng H, Kim B-H, Grishin NV. Analysis of CASP8 targets, predictions and assessment methods. Database, Vol. 209 Article ID bap 003.
13. Eastwood MP, Hardin C, Luthey-Schulten Z, Wolynes PG. evaluating protein structure prediction schemes using energy landscape theory. IBM J Res Dev 2001;45.
14. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. Optimal protein-folding codes from spin-glass theory. Proc Natl Acad Sci USA 1992;89:4918–4922.
15. Kinch LN, Qi Y, Hubbard TJ, Grishin NV. CASP5 target classification. Proteins 2003;53(Suppl 6):340–351.
16. Zong C, Papoian GA, Ulander J, Wolynes PG. Role of topology, nonadditivity, and water-mediated interactions in predicting the structures of alpha/beta proteins; proteins. J Am Chem Soc 2006; 128:5168–5176.
17. Ortiz AR, Strauss CE, Olmea O MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 2002;11:2606–2621.
18. 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.
19. Prlic A, Down TA, Hubbard TJ. Adding some SPICE to DAS. Bioinformatics 2005;21(Suppl 2):ii40–ii41.
20. DeLano WL The PyMOL Molecular Graphics System: DeLano Scientific, San Carlos, CA. Available at: http://www.pymol.org.
21. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
22. Onuchic JN, Wolynes PG. Theory of protein folding. Curr Opin Struct Biol 2004;14:70–75.
23. Shakhnovich E. Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. Chem Rev 2006;106: 1559–1588.
24. Dill KA, Ozkan SB, Shell MS, Weikl TR. The protein folding problem. Annu Rev Bipohys Biomol Struct 2008;37:289–-316.
25. Plotkin SS, Onuchic JN. Understanding protein folding with energy landscape theory. Part I: basic concepts. Quart Rev Biophys 2002; 35:111–167.
26. Oliveberg M, Wolynes PG. The experimental survey of protein-folding energy landscapes. Quart Rev Biophys 2005;38:245–288.
27. Fersht AR, Daggett V. Protein folding and unfolding at atomic resolution. Cell 2002;108:573–582.
28. Service RF. Problem solved* (*sort of). Science 2008;321:784–786.
29. Aloy P, Stark A, Hadley C, Russell RB. Predictions without templates: new folds, secondary structure, and contacts in CASP5. Proteins 2003;53(Suppl 6):436–456.
30. Noivirt-Brik O, Prilusky J, Sussman JL. Assessment of disorder predictions in CASP8. Proteins, in press.