WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

New Folds: Assessment

Assessment of CASP7 structure predictions for template free targets

Ralf Jauch, Hock Chuan Yeo, Prasanna R. Kolatkar, and Neil D. Clarke*

Computational and Systems Biology, Genome Institute of Singapore, Singapore

ABSTRACT

In CASP7, protein structure prediction targets that lacked substantial similarity to a protein in the PDB at the time of assessment were considered to be free modeling targets (FM). We assessed predictions for 14 FM targets as well as four other targets that were deemed to be on the borderline between FM targets and template based modeling targets (TBM/FM). GDT_TS was used as one measure of model quality. Model quality was also assessed by visual inspection. Visual inspection was performed by three independent assessors who were blinded to GDT_TS scores and other quantitative measures of model quality. The best models by visual inspection tended to rank among the top few percent by GDT_TS, but were typically not the highest scoring models. Thus, visual inspection remains an essential component of assessment for FM targets. Overall, group TS020 (Baker) performed best, but success on individual targets was widely distributed among many groups. Among these other groups, TS024 and TS025 (Zhang and Zhang server) performed notably well without exceptionally large computing resources. This should be considered encouraging for future CASPs. There was a sense of progress in template FM relative to CASP6, but we were unable to demonstrate this progress objectively.

Proteins 2007; 69(Suppl 8):57–67.

Key words: CASP7; free modeling; structure prediction; assessment.

INTRODUCTION

The last several editions of the biennial CASP experiment have shown progress in the prediction of novel protein structures, certainly when compared with the first couple.[1–3] This progress has come largely from biased sampling of structural fragments from the PDB to assemble initial models, an idea that is now 10-years old.[4,5] Biased sampling of structural fragments restricts dramatically the conformational search space required to construct plausible models. However, *de novo* structure prediction still requires substantial computing resources and is still an extraordinarily challenging problem. It remains unclear, for example, what the optimal search strategies are for fragment selection, fragment assembly, and structure refinement, and there is continuing research to define appropriate energy functions for picking the best models from among the large numbers of candidate structures that are generated.

It is probably fair to say that the best *de novo* predictions from the last few CASPs have been remarkable by the expectations of most observers. *De novo* methods have, in some cases, rivaled or even exceeded the performance of template based methods when the available template is only distantly related to the target. In general, though, *de novo* model quality has remained poor compared with models that are based on templates.

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, and any quantitative measure of similarity, is an implicit (and imperfect) statement about what is considered to be important in a structure prediction. This is a particularly vexing and subjective problem when it comes to the assessment of structure predictions for novel folds, which are generally

The authors state no conflict of interest.

*Correspondence to: Neil D. Clarke, Computational and Systems Biology Group, Genome Institute of Singapore, 60 Biopolis St, Singapore 138672. E-mail: clarken@gis.a-star.edu.sg

Received 28 February 2007; Revised 10 July 2007; Accepted 13 July 2007
Published online 25 September 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21771

much less accurate than template-based models. What are the attributes of a structure prediction that make it good, or perhaps even useful? Is it the overall chain topology? If so, how can we define and quantify topological similarity? Is the accuracy of the secondary structural elements relevant? How much weight should be attached to partial structures that are modeled with higher accuracy compared with the quality of the overall topology?

GDT_TS[6] is a generally accepted measure of backbone similarity for evaluating template-based models and has been used over the last several CASPs. It has also been used to assess new fold predictions, though it is less clear for these (generally) poorer models how well the GDT_TS score correlates with what a structural biologist would consider to be a good model. For that matter, when it comes to new-fold predictions, it is not clear how well structural biologists agree amongst themselves on what the best models are.

As part of our assessment of free modeling (FM) in CASP7, we sought to address some of the issues regarding assessment itself. We evaluated the consistency with which independent individuals identified good models by subjective criteria, and we compared the subjective assessments with two objective measures of structural similarity. These two measures were GDT_TS, which has been extensively used in previous CASPs, and a contact map similarity measure that was devised for this study. We conclude that GDT_TS is useful for prioritizing models for visual inspection but that visual assessors often agree that other models are better than the best GDT_TS model. The contact map similarity metric is typically less predictive of the best visual models than is GDT_TS.

METHODS

Targets

Classification of all CASP targets resulted in a set of 15 free modeling targets (FM) and four targets that were deemed to be on the border between FM and template based modeling (TBM/FM) (see article elsewhere in this issue for the classification of targets). Eighteen of these 19 targets were used in the analyses described here. Regrettably, we failed to notice that T0316_2 had been classified as an FM target until our analyses were complete. A retrospective analysis of predictions for T0316_2 indicated that the inclusion of this target would not have affected our conclusions (not shown).

Objective scoring of target-model similarity

We used two objective measures of model-to-target similarity to prioritize models for visual assessment. One was GDT_TS, which was calculated by the CASP Prediction Center. The second measure we call Contact Map Overlap (CMO). Briefly, lists of contacts were obtained for target structures and for models. For the models, only residues that were also found in the target structure were used to construct contact lists. For each target and model, six different contact lists were generated. Specifically, we calculated contacts using distance thresholds of 8 Å and 12 Å, and for each of these thresholds, we considered residue pairs in which the residues were at least 6, 12, or 24 residues apart. The choice of these parameters was based in part on contact prediction assessments in previous CASPs,[7] and in part on an empirical analysis using targets and models from CASP6 that was conducted before any CASP7 data was examined. Contacts here were defined between pseudo-Cα positions. For a residue, i, the pseudo Cα position is the Euclidean average of Cα coordinates for residues $i-1$, i, and $i+1$. Residues lacking a preceding or succeeding residue were not used in the calculation of contact maps. This averaging was done with the idea that a side chain-based contact definition, or even a Cα-based position that used only the coordinates of the residue itself, would overly penalize a model that had a very good overall chain topology, but which had secondary structure elements that were out of phase (i.e., rotated) with respect to the target structure. For each of the six contact map definitions given above, the similarity, S, of target and model contact maps was defined as $S = \frac{|T_C \cap M_C|}{(M_T + T_T)/2}$ where $|T_C \cap M_C|$ is the number of contacts in common between the target and model, M_T is the total number of residue pairs in the model that meet the sequence separation criterion, and T_T is the total number of residue pairs in the target that meet the sequence separation criterion. For each contact map definition, the score for each model was converted to a Z-score based on the mean and standard deviation of the contact map similarity scores. The overall score for a model was then the average of the six Z-scores.

Ranking and statistical tests based on GDT_TS

Groups were ranked based on the GDT_TS scores for their best models. For all pairs of groups, targets predicted in common by the groups were used as the basis of comparison. A group was said to have done better than another group in the pairwise evaluation if it had a higher GDT_TS score for more than half the targets. Groups were then ranked according to the number of other groups that they beat by this criterion. To assess the statistical significance of the pairwise comparisons, raw GDT_TS scores for each target were converted to Z-scores (number of standard deviations away from the mean). The Student t test was then used to determine whether there was a significant difference in the mean Z-scores for targets predicted in common between two groups.

Visual assessment

Molecular graphics

Visual inspection was performed using PyMOL.[8] To facilitate the analysis, we generated a PyMOL script for each target-model pair and all visual assessments started

DOI 10.1002/prot

with the running of that script. Molecules were displayed using the "cartoon tube" representation to avoid biases due to the visual appeal of secondary structures drawn in the usual cartoon style of protein structures. In particular, we wished to avoid being biased against modelers whose overall fold might be good but who did not impose strong secondary structure constraints and whose secondary structures might therefore not be rendered as secondary structures by PyMOL. Target and model structures were superimposed using PyMOLs "align" command, though it was often found easier to analyze if one molecule were subsequently moved with respect to the other. To aid the visual identification of regions that correspond to one another in targets and models, gradient coloring was applied such that corresponding residues in the target and model were colored the same. PyMOL scripts from Robert Campbell were found helpful for this purpose (http://adelie.biochem.queensu.ca/~rlc/work/pymol/).

Selection of models for visual assessment

Visual inspections were performed independently by three individuals (RJ, PRK, and NDC, henceforth referred to as the assessors). For each target, each assessor was given a list of 75 models to assess. Fifty of these models were in common to all assessors, and were selected on the basis of GDT_TS and CMO scores as follows. First, the top 25 models by GDT_TS and the top 25 models by CMO were selected. As there were always models in common in these sets, their union had fewer than 50 models. To bring the set of prioritized models up to 50, the GDT_TS and CMO scores for all models were first normalized by conversion to a *Z*-score, and then the GDT_TS *Z*-score and the CMO *Z*-score for each model were averaged and used to rank all models that had not already been selected based on GDT_TS or CMO scores alone. This list was used to bring up to 50 the total number of models selected by objective criteria. Each assessor also assessed 25 randomly selected models that were uniquely assigned to him. The lists of models given to each assessor were randomly sorted so that the assessor was unaware which models had been selected by objective criteria and which had been selected randomly.

Identifying sets of top models by visual assessment

We devised a multistage, multiassessor procedure to minimize the chances that we would miss a prediction group who had done well on a particular target. In the first round, each assessor independently short-listed a set of five groups, and the associated model(s) that they considered best for each target. The union of these selected models was then generated and sent back to the assessors for a second round of independent (re-) assessment. This approach ensured that individual assessors were exposed to the top selection of their colleagues in a round of blinded reconsideration. Each assessor produced a list of the top three groups and the associated model(s). Where possible, assessors also ranked the groups within these lists. In a small number of cases, an assessor felt that the top one or two models were so clearly better than several other models that bringing the total list up to three would be arbitrary. In the third and final round of visual assessment, the assessors met to discuss their evaluations. During these discussions we did not allow ourselves to switch our original choices to other models. However, in some cases, individual assessors became convinced by the discussion that one of their "final-three" selections was less accurate than other models and chose to withdraw their selection. Out of a total of 162 possible votes cast (3 assessors $\times$ 3 choices $\times$ 18 models), 136 were left at the end. If a predictor group had multiple models left in contention at this stage, the model with the highest GDT_TS score was chosen to be representative of that group's predictions.

RESULTS

Performance by group

A total of 36 groups produced at least one model that was short listed after the three rounds of visual assessment (Table I). Group TS020 (Baker) clearly did best by this criterion as they received a total of 18 votes for the 18 targets that were evaluated. This is one-third of the maximum as there were three assessors who could have selected a TS020 model for each target. Four groups received a total of 7 votes (TS004, TS013, TS025, TS197) and four received a total of six votes (TS024, TS026, TS125, TS178). By chance, we would not have expected any group to receive six votes and only one group to have five.

Groups were also compared using the GDT_TS scores of their best models (Fig. 1, see Methods). By this criterion, TS020 and TS024 rank at the top and cannot be statistically distinguished from each other by pairwise comparison. However, TS020 is significantly better than a few other high-scoring groups that TS024 itself cannot be distinguished from (note the comparisons to TS025, TS050, and TS004; Fig. 1). In that sense, TS020 outperformed TS024. By the same criterion, TS024 in turn was better than all other lower-ranked groups, including the third ranked group, TS025. It is noteworthy, however, that even the best groups cannot be distinguished with confidence from every group. For example, neither TS020 nor TS024 can be distinguished from TS033 (ranked 5th) because the latter group, though it did well on the targets it predicted, had substantially fewer predictions, making it more difficult to establish statistical significance.

Many of the top-scoring groups by GDT_TS [Fig. 1(B)] were also among the groups identified most frequently in the visual assessment. As already noted, group TS020 performed best by both criteria. Groups TS024 and TS025 (Zhang and Zhang_server, respectively) were second and third best by GDT_TS and were selected a combined total of 13 times in the visual assessment. Other groups found in

Table I
Models Selected as Being Among the Top 3 by 2 or More of the 3 Visual Inspectors

Target	Model	# Top 3 selections	GDT_TS rank	CMO rank
T0287	TS024_2	2	3	17
T0287	TS111_2	2	4	16
T0287	TS414_4	2	50	55
T0296	TS125_3	2	6	7
T0300	TS033_2	3	1	21
T0300	TS047_1	3	13	76
T0300	*TS651_1*	3	15	78
T0304	TS564_4	2	6	2
T0304	TS020_3	2	9	1
T0304	TS276_2	2	14	80
T0304	TS338_3	2	62	3
T0307	TS178_2	3	10	14
T0307	TS337_4	2	6	134
T0307	TS025_3	2	8	116
T0309	TS415_5	3	19	44
T0309	TS277_5	3	18	19
T0314	TS178_3	3	5	27
T0314	TS025_5	2	1	3
T0319	TS020_3	3	1	1
T0319	TS197_2	2	16	85
T0321_D2	*TS020_1_2*	3	24	83
T0321_D2	TS213_1	2	14	89
T0347_D2	TS020_4	3	1	76
T0347_D2	TS249_4	2	10	52
T0347_D2	TS013_2	2	3	83
T0348	TS026_5	3	1	42
T0348	TS027_3	2	1	41
T0348	TS211_1	2	10	22
T0348	TS004_4	2	12	24
T0350	TS004_2	2	3	6
T0350	TS050_2	2	6	4
T0350	TS710_5	2	11	12
T0353	TS013_5	3	20	30
T0353	TS020_3	3	24	17
T0353	TS004_4	2	10	13
T0356_D1	TS020_1	3	1	323
T0356_D1	TS013_1_1	2	8	363
T0356_D3	TS197_2	3	3	31
T0356_D3	TS047_3	2	22	19
T0356_D3	TS034_3	2	27	32
T0361	TS125_3	3	34	21
T0382	TS026_4	3	2	58
T0382	TS024_3	2	1	14
T0386_D2	TS010_5	2	1	99
T0386_D2	TS469_4	2	25	NA

The number of assessors listing the model is indicated in the column label as "#top 3 selections." The "GDT_TS rank" and "CMO rank" of the model show the ranking of the model by GDT_TS and CMO, respectively, compared with all models for that target. The models whose names are italicized were identified by all three assessors as the single best model for that target, not just one of the top three. Because of an uncharacterized bug in the CMO software, model TS469_4 of target T0386_D2 could not be scored and is therefore given a rank of 'NA.'

the top 20 by GDT_TS and selected at least five times by visual assessment were TS004 (ROBETTA), TS013 (Jones-UCL), TS047 (Pmodeller6), and TS125 (Tasser).

Comparison of visual and GDT_TS based assessments

In general, the top models identified by visual assessment rank reasonably well by GDT_TS. Nearly all the identified models rank in the top 25 by GDT_TS, out of roughly 500 models for each target [Fig. 2(A)]. However, the top model by GDT_TS was not always identified as one of the top models visually [Fig. 2(B)]. To verify that we did not accidentally miss the best GDT_TS models, five targets were reassessed a few weeks after the visual assessment was completed. For these targets, the top-scoring GDT_TS model was not identified by any of the visual assessors and all three assessors independently identified

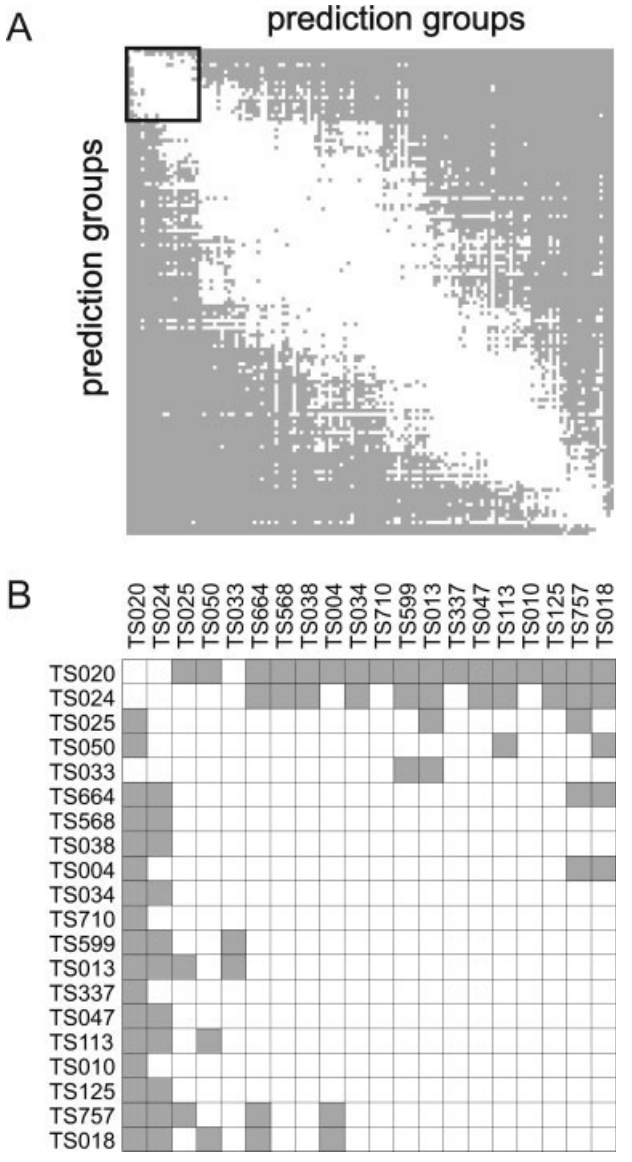

Figure 1
Ranking of prediction groups by pairwise comparison of GDT_TS scores. ***A****: All groups were ranked by pairwise comparison of the best models from each group using the targets predicted in common by the two groups (Methods). Pairs of groups whose normalized GDT_TS scores are significantly different from one another are shown in gray (P ≤ 0.05).* ***B****: Inset of panel A, showing top-ranked groups. This plot corresponds to the square-enclosed region in the upper-left corner of panel A.*

DOI 10.1002/prot

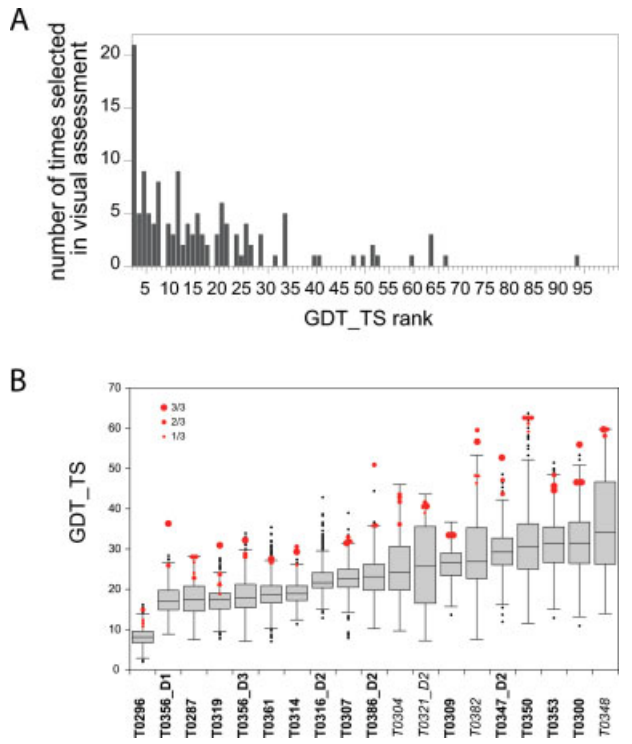

Figure 2

GDT_TS scores for top models identified by visual assessment. ***A****: The number of times a model was selected by visual assessment and had the indicated GDT_TS rank. One model that was selected by a single assessor ranked 233 and is not shown on this plot.* ***B****: Box plots show the distribution of GDT_TS values for all 19 FM and TBM/FM targets. Italicized target names are for the four TBM/FM targets. Red circles indicate the GDT_TS scores for models identified as the top models by visual assessment. The size of the red circles corresponds to the number of assessors who picked that model. Target T0316_D2 was not visually assessed.*

another model with a substantially lower GDT_TS score. For each target, the model pairs were randomly renamed "A" and "B" and assessors were asked for their preference. Out of 15 reassessments (5 targets × 3 assessors), 14 again preferred the model that scored lower by GDT_TS.

Individual targets

We visually assessed 14 free-modeling (FM) targets and four targets from the borderline between FM and template based modeling (TBM/FM). We briefly highlight the challenges presented by each target, and the nature of the successes and failures in predicting each target. Targets are discussed in numerical order for ease of reference. In the interests of space, only a subset of the FM targets and their most instructive models could be selected for illustrations (Figs. 3–9).

T0287. This target is a bundle of 10 differently sized helices arranged in an irregular up-and-down topology. Difficulties were encountered predicting helical boundaries, kinks and turns, especially at the C-terminus. Top models nicely predicted the N-terminal and central part of the protein but introduced strikingly similar errors at the C-terminus. There were possible templates that had some helical elements in common (e.g., 1v55N), but they are extremely remote and it does not appear that any predictor was able to use these structures explicitly.

T0296. The size and architecture of this protein made it arguably the most difficult target of CASP7. It is a α/β protein with an incompletely closed barrel at the center surrounded by irregularly arranged helices. The topology vaguely resembles that of a TIM barrel, except the chain crosses back and reverses direction part way through the structure. The section starting near this switch was evidently difficult to model, as many groups had reasonable α/β structures at the N-terminus but only very limited resemblance to the target overall.

T0300. This protein consists of three α-helices and an extended strand that forms a two-stranded antiparallel

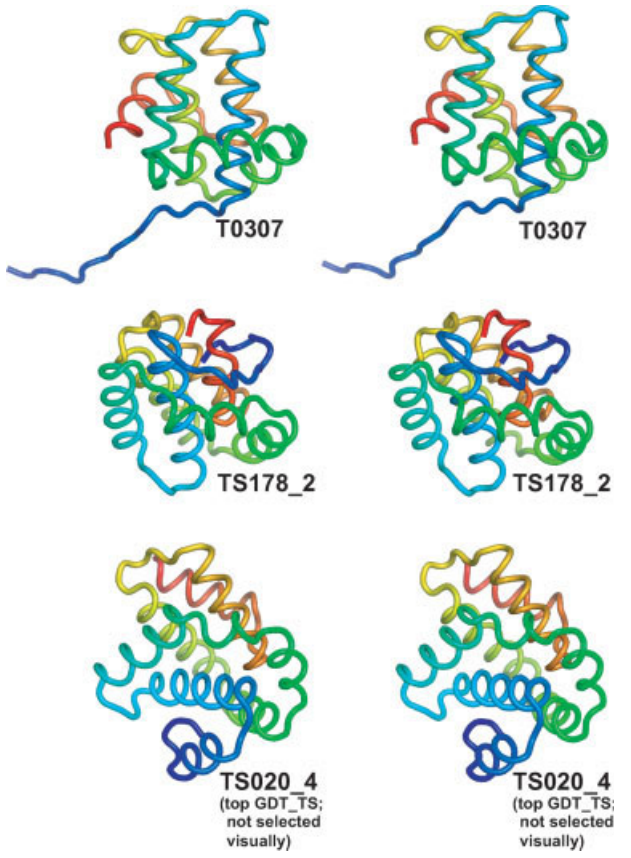

Figure 3

An example where the best GDT_TS model was not considered a good model by visual inspection. TS178_2 is one of three models for target T0307 that were identified as best by visual inspection (Table I). Secondary structure elements are substantially correct, and some supersecondary structures are accurate as well. In contrast, TS020_4, the best model by GDT_TS, adopts a kind of superhelical structure. No assessor selected this model. In this figure and in others, structures are drawn in stereo with PyMOL.[8] *Color gradients run from blue at the N-terminus to red at the C-terminus.*

sheet with another monomer in the asymmetric unit. The monomers differ substantially in the orientation of one of the terminal helices due to a helix-swap packing arrangement with a crystallographic symmetry mate of the dimer. The overall structure is clearly a tetramer (dimer-of-dimers) in the crystal. Many groups managed to predict the monomer fold and correctly assigned the structural elements even though remote templates align only to fragments of the target (e.g., the long central helix). Despite the similarity of many of these models to each other, the three assessors each identified the same set of three models as the best predictions (Table I). A number of groups attempted to model oligomeric structures as they had been told that the target was a dimer. Some of these oligomer models were quite plausible, but none can be considered accurate. It does not appear that oligomer modeling helped to predict the structure of the monomer.

T0304. This TBM/FM target consists of a twisted sheet of 21543 topology wrapping around a kinked helix located between strand 2 and 3. Many models were quite good beyond the first 15 amino acids even though there is no evidence for explicit use of the available remote templates (e.g., 2gnx_A). The most common inaccuracies seen in visually selected models were errors in the sheet topology, failures to predict the curvature of the sheet, and the kink of the helix. Model TS276_2 accurately predicts most of the model's key features but invents a long protrusion and a short helix at the N-terminus at the expense of severely shortened strands 1 and 2. Model TS020_3 exhibits the helical kink and a largely correct sheet topology. It lacks, however, the correct sheet curvature and places the C-terminal part of strand 2 as a 6th strand resuming the sheet. The latter inaccuracies are also seen in models TS338_3 and TS564_4, which were also selected by visual assessment.

T0307. This protein adopts a compact all-α structure consisting of seven helices (see Fig. 3). There was a plausible template based on structural similarity criteria (1gn3n_C) but it lacked two of the internal helices found in T0307 and it does not appear to have been used. Many groups recognized the helical nature of the fold but predicted a superhelical arrangement of the helices. Other groups managed to predict a helical bundle, but with individual helices oriented incorrectly. Three models were consensus winners by visual assessment (Table I). These rank within the top 10 by GDT_TS but, interestingly, the model ranked highest by GDT_TS has an incorrect superhelix-like arrangement of helices (see Fig. 3).

T0309. This was a difficult target as it forms an octamer in which two strands from each subunit interdigitate with the strands of neighboring subunits to form a 16 stranded β-barrel, with additional elements adorning the outside of the barrel (see Fig. 4). Visually selected models succeeded in predicting the helix and traced the main-chain surprisingly well. It was not possible to model the β-structures accurately without anticipating the oligomeric structure, but the best models did form intramolecular sheet structures that include many of the sheet-forming residues found in the target (see Fig. 4).

T0314. This protein contains two 3-stranded sheets and seven helices in an irregular arrangement (see Fig. 5). Despite being a small, monomeric target, this was evidently a difficult fold to predict. The top visual models ranked high by GDT_TS and correctly modeled some supersecondary structural elements. However, even the best models contained substantial errors. TS025_5, which ranked first by GDT_TS and was also independently picked among the top three models by two of the three assessors, illustrates the limits of prediction for this target (see Fig. 5).

T0319. This was a difficult target, as indicated by low GDT_TS scores and by visual criteria. By both measures, though, model TS020_3 stood out. There are two subdomains to this target (see Fig. 6). One is a three-stranded sheet with N- and C-termini packed around it in irregular structures; the other, made up the central portion of the sequence, consists of an irregular three helix-bundle and a long loop between helix 2 and 3 that packs against helix 3. Model TS020_3 captures the essential features of these subdomains, though it fails to predict accurately the relative orientation of the subdomains. This model was selected by all three assessors and also had the highest GDT_TS score. TS197_2 was deemed the second best model but individual secondary structure elements, although present, are arranged incorrectly and there is no clear separation of the structure into distinct subdomains (see Fig. 6).

T0321_D2. This TBM/FM target had a template covering the central α/β region of the structure and, not surprisingly, many groups got that part of the structure correct. However, the C-terminal strand and the N-terminal ~30 amino acids were true free-modeling problems. TS020_1_2 was independently chosen by all three assessors as the single best model for this target. Interestingly, this model ranks only 24th by GDT_TS. Among other things, TS020_1_2 correctly predicts the antiparallel orientation of β-strand 6, which caused difficulties for other groups.

T0347_D2. This short domain consists of just three helices, packed in a way that is at least reminiscent of many other structures [Fig. 7(A)]. However, there was no template by the criteria that were used to classify target structures (see accompanying paper on target classification in this issue). Many groups had very good models for this target, with only small deviations from the target structure. Even by these standards, though, TS020_4 stood out by GDT_TS [Fig. 2(B)] and by visual inspection (the only model picked by all three assessors).

T0348. This target had a template covering the three central strands, leaving the N- and C-terminal regions as the

 DOI 10.1002/prot

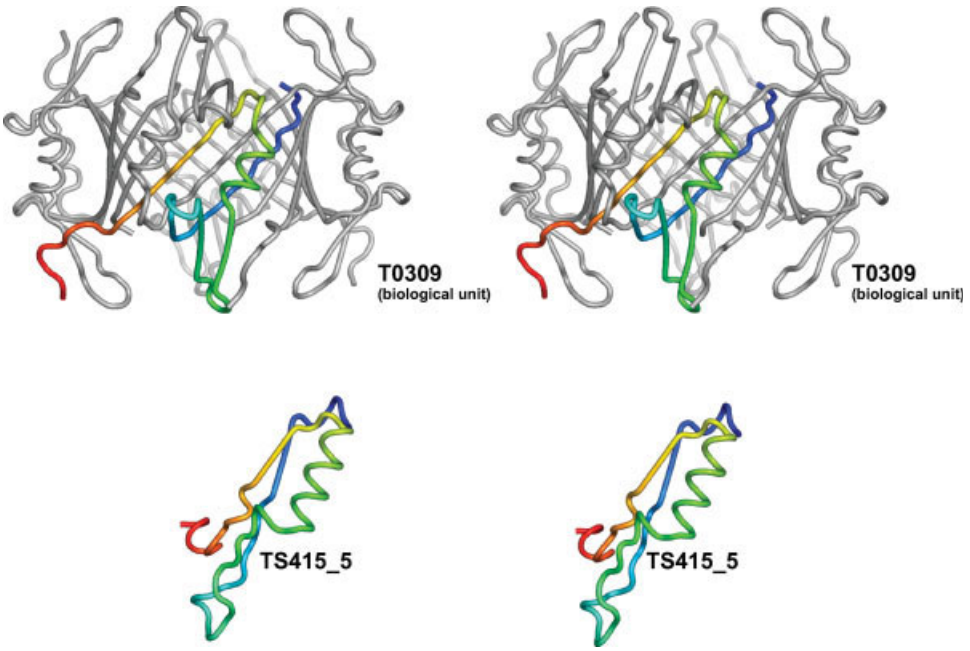

Figure 4

An oligomeric target that complicates accurate prediction of the monomer. Despite the oligomeric structure of T0309, TS415_5, and other good models correctly predict many of the secondary structure elements and their orientation.

only true FM portion of the structure. Prediction of this target was complicated by the fact that the protein dimerizes along a noncrystallographic twofold axis mediated partly by a swapped N-terminal strand that extends the sheet of the adjacent molecule. However, in some models the N-terminal strand participates in an intramolecular sheet that is analogous to the strand-swapped sheet observed in the crystal structure. A distinguishing feature of the best models is the correct prediction of a C-terminal helix that packs against the sheet opposite of the dimer interface.

T0350. This target consist of three helices, packed almost as a sheet, lying on a twisted three-stranded sheet [Fig. 7(B)]. Although we were not able to identify a template for this target, many groups submitted predictions with an astonishing degree of accuracy. Indeed, the similarity of many excellent models made it difficult to nominate the three top models visually. Six models were selected by visual inspection, all of which also rank high by GDT_TS (Table I). Some of those originate from publicly accessible servers (i.e., TS004_2) which could in principle have been used by other groups as well. The secondary structures of top models superimpose very closely and only the N-terminal loop and the loop connecting helix 1 and 2 exhibit noteworthy differences. It is not clear why this target was so easy. The β-sheet is much simpler than in other targets, but the connectivity of the helices before and after the sheet does not seem trivial. Perhaps getting the β-sheet structure right imposes strong constraints on the arrangement of the helices.

T0353. This target is an α/β protein with two helices and an antiparallel sheet with 3124 topology (see Fig. 8). No model got the topology of this structure exactly right. The

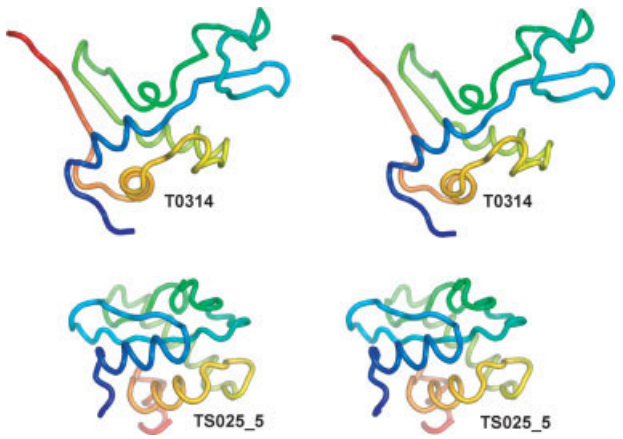

Figure 5

An example of a target that lacked good models despite being small and monomeric. TS025_5 was the top-ranked model for target T0314 according to GDT_TS and it was also one of the best models by visual inspection. In common with some other top models for this target, it predicts too many helices and displays substantial errors with regard to the overall topology.

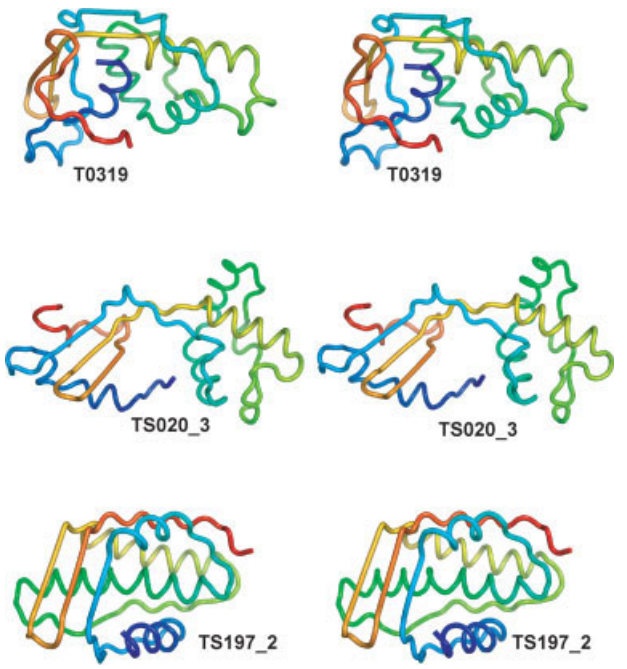

Figure 6

An unusually good model for a difficult target. Model TS020_3 was identified as the single best model for target T0319 by all three assessors, and was ranked first by GDT_TS. TS197_2, ranked second best by visual inspection, is shown along with TS020_3 to highlight the features that made the latter model stand out. These features include the topologies of the helical and β-sheet subdomains, and the separation of these subdomains. The orientation of the subdomains, however, is incorrect.

two models favored by all three assessors (TS020_3 and TS013_5) exhibit a near-perfect N-terminal region but encountered similar problems at the C-terminus. Both models align strand 4 in the wrong orientation and predict an additional helix instead of the extension of strand 3. Other relatively good predictions did not model strand 3 at all.

T0356_D1. T0356_D1 is a noncontinuous target consisting of residues 7–96 and 314–347 of the parent structure. Several groups managed to predict parts of the N-terminal fragment, but generally failed to arrange the secondary structural elements correctly. Model TS013_1_1 represents an excellent solution for the N-terminus except that helices 3 and 4 point in the wrong directions. The single biggest challenge was the correct docking of the two noncontiguous sections of this target. To succeed, predictors had to model a much larger protein in a way that would place these segments accordingly. Model TS020_1 got closest in this respect and also exhibits an exceptional prediction of the N-terminal region.

T0356_D3. The C-terminus of target T0356 (residues 348–467) constitutes the second FM domain of this protein. It consists of a five-stranded sheet with two helices inserted between strands. The sheet has a 1,2354 topology with parallel as well as antiparallel strands. No group predicted this sheet correctly, with C-terminal strands 4 and 5 being the most problematic part of the structure. The top predictions achieved a decent degree of accuracy with respect to the domain's N-terminus. By visual assessment, the top model by GDT_TS is less accurate than the models selected subjectively. This was confirmed in a blind validation test (see above).

T0361. This all-α protein consists of an unusual pair of roughly coplanar helical layers that are rotated with respect to each other by about 45°. No group predicted this arrangement accurately, but credit was given to models that roughly resemble the ordered helical arrangement. Model TS125_3 was the only model that was picked by all three assessors, but it exhibits a significantly more compact, bundle-like shape than the target. Distinguishing features of this model are the arrangement of the N- and C-termini and the overall trace of the chain. It is noteworthy that none of the models chosen by the assessors ranks within the top-20 using either of the objective scoring functions that were used to prioritize models for visual assessment.

T0382. This is an all-α protein consisting of six helices. The helices appear to be arranged as three pairs of two helices forming the start of a HEAT repeat-like superhelical structure. A number of reasonable predictions were made for this target although an unambiguous

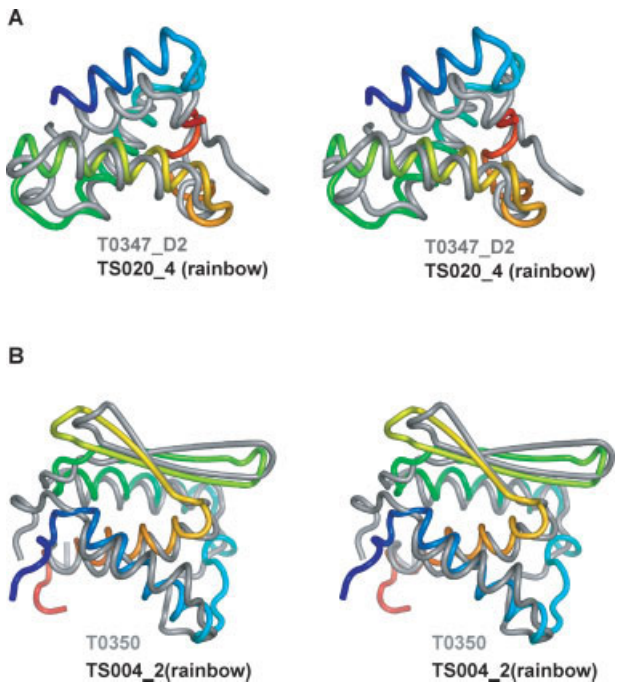

Figure 7

*Superposition of model and target for two very well predicted FM targets. **A**: Target T0347_D2 and model TS020_4. **B**: Target T0350 and model TS004_2.*

DOI 10.1002/prot

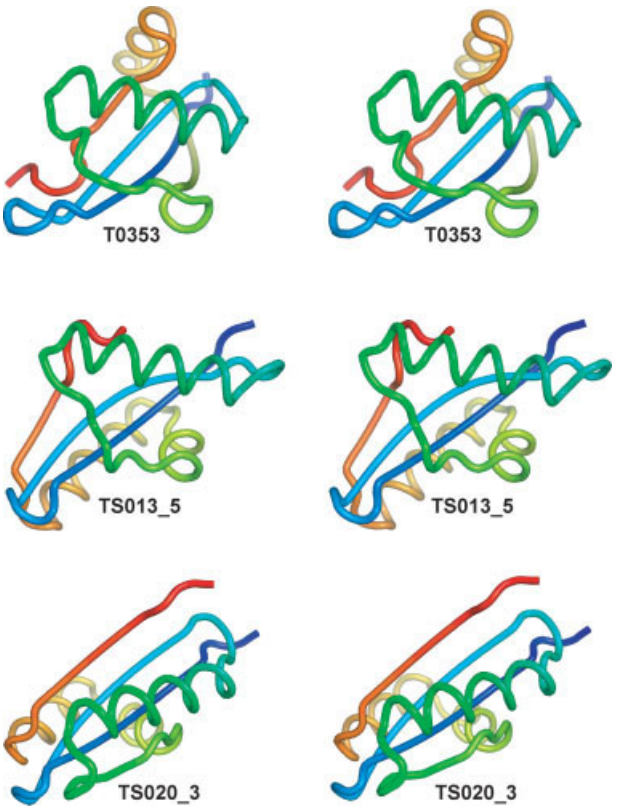

Figure 8

Models for target T0353 have similar errors at the C-terminus. Models TS020_3 and TS013_5 were the two models favored by all three assessors. Both models are quite accurate in the N-terminal region, but make similar errors at the C-terminus. Note the incorrect sheet topology.

template could not be found. The top-two models by visual assessment were also the top-two models by GDT_TS (Table I).

T0386_D2. The C-terminal domain (residues 219–299) of target T0386 consists mostly of a β-sheet with a complicated 1,4325 topology, and an N-terminal helix that packs against the sheet (see Fig. 9). The sheet posed a serious challenge to predictors. Model, TS010_5, however, stood out by visual inspection and by GDT_TS (see Fig. 9). It exhibits a roughly correct strand topology but failed to predict the mutual packing of helix and sheet.

DISCUSSION

Quality of predictions

For a couple of targets, a number of models were extremely good and could be superimposed on the targets with no large scale deviations (see Fig. 7). For T0347_D2, this might be attributed to the fact that it is a simple, three-helix structure. Even so, the accuracy of the best model, TS020_4, is remarkable. This may be a case where the computationally expensive refinement procedures employed by Baker and colleagues made a real difference. For target T0350, the quality of the models is unexpected. This target consists of a three-stranded antiparallel sheet flanked by two helices at one end and one at the other [Fig. 7(B)]. While the sheet topology is simple, this is clearly a much more complicated structure than T0347_D2 [Fig 7(A)]. Nevertheless, a number of groups did remarkably well. Note the number of GDT_TS outliers for this target and the fact that the top GDT_TS scores for this target are higher than for any other target; Figure 2(B). To some extent, the number of groups performing well could be due to server-based predictions being used as we found examples of models with identical backbone coordinates. But this begs the question of how any group did as well as they did. Perhaps the overall fold is more tightly constrained by the secondary structure elements than is obvious from looking at the structure.

Other targets were notable because a single model succeeded in capturing important structural features that were lacking in other models. Good examples are model TS020_3 for target T0319 and model TS010_5 for target T0386_2. These models are nowhere near as accurate as the models for T0347_D2 and T0350 described above, but they succeeded in getting some overall characteristics of the fold correct for targets that were clearly quite difficult. Most targets and their models, though, were less satisfying than this. The best predictions were often quite good at the secondary structure level, and for some super-secondary structural motifs, but the chain typically went off in the wrong direction at one or more points.

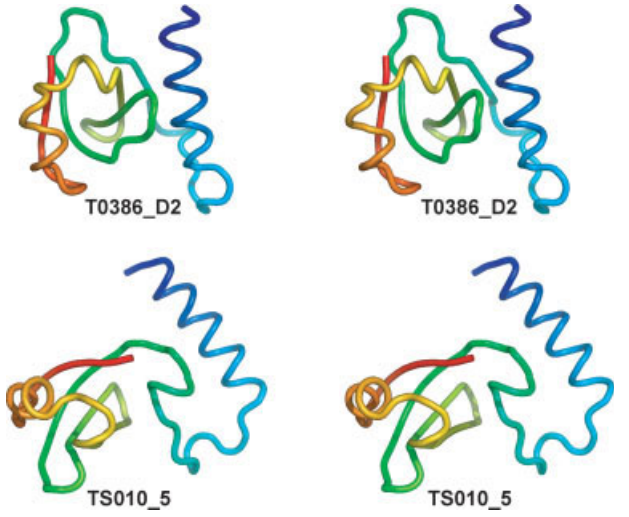

Figure 9

Correct prediction of sheet topology for a difficult target. Target T0386_D2 has a complicated β-sheet topology. Model TS010_5 stood out from other models in getting the strand order and orientation correct.

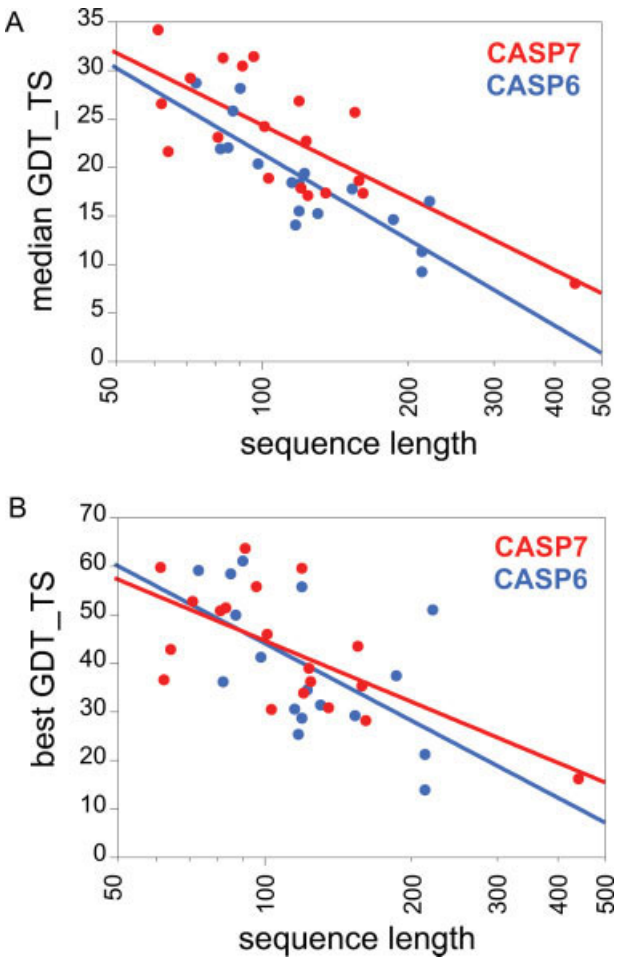

Figure 10

*GDT_TS scores for targets in CASP7 and CASP6. **A**: Median GDT_TS scores as a function of target size. FM and TBM/FM targets from CASP7 are shown in red; the targets from CASP6 that are most analogous to these (NF and FR/A) are shown in blue. **B**: Same as panel A except the best GDT_TS value for the target is plotted rather than the median value.*

Progress in *de novo* structure prediction?

One presumes that progress has been made since CASP6, if only because the Baker group, which has topped the new fold/FM category for several CASPs running, continues to develop their methods and to use ever-increasing computing power. It was not possible, though, to establish objectively that progress had indeed been made. The main difficulty in comparing results across CASP experiments is that the targets are different. Target difficulty depends on many factors: size, secondary structure composition and arrangement, number and diversity of sequence homologs, extent to which remote or partial templates exist, and so on.

Of the differences in targets that affect target difficulty, size is the easiest to control for. Figure 10(A) shows that the median GDT_TS score for CASP7 targets tends to be higher, for a given target size, than the median GDT_TS score for CASP6 targets. This suggests that a number of prediction groups did better in CASP7 than in CASP6. However, when the best GDT_TS scores are used rather than the median values, no improvement in CASP7 over CASP6 is observed [Fig. 10(B)]. Thus, while there may have been more groups that did reasonably well, there is no evidence that the best predictions were better than in CASP6. Furthermore, the encouraging trend seen for the median GDT_TS values disappears when targets that are on the borderline of being template-based are removed from the analysis (these were called TBM/FM in CASP7 and FR/A in CASP6; data not shown). Thus, the evidence for improvement in CASP7 is limited to median performance and even this depends on targets for which there were remote and/or partial templates.

There was a perception among some predictors that targets were harder in CASP7 than in CASP6, and that, as a result, GDT_TS scores may have failed to capture an improvement in prediction quality. The greater difficulty was ascribed to differences in the number and diversity of sequence homologues for the target sequences. There is some support for this perception as there were more CASP7 FM targets that lacked homologues altogether (4) than was the case for CASP6 NF targets (1) (Michael Tress, personal communications). It is not clear how to extend this analysis to targets that have homologues, unfortunately, because the relationship between the num-

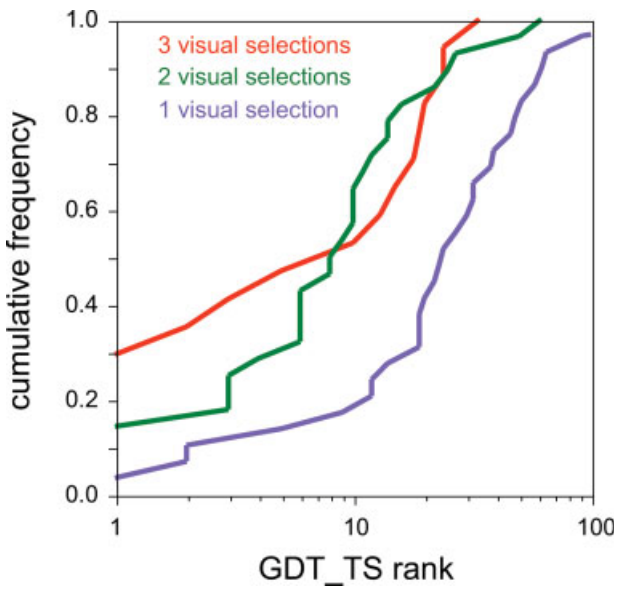

Figure 11

Cumulative probability of being identified as a top model by visual assessment as a function of GDT_TS rank. Models were classified depending on whether they were picked in the visual assessment by one, two, or three assessors. In general, models that are picked more frequently in the visual assessment tend to rank higher by GDT_TS.

DOI 10.1002/prot

ber and diversity of homologues and the difficulty of the target is not obvious. For future CASPs, it may be worthwhile to document, at the time of prediction, the number of sequence homologous detectable by some proscribed method (e.g., PSI-BLAST with default parameters), and a measure of the sequence similarity to the target.

Current methods seem capable of producing quite accurate models for certain targets (e.g., T0350). More such high-accuracy successes can be expected in CASP8 simply due to the increasing size of fragment libraries and growth in computational resources. On a more discouraging note, it is evidently still not possible for the structure prediction community to get the overall fold correct for the majority of targets. For a subset of targets, there may be extenuating circumstances for this failure, such as large size or unusually complicated interactions among domains or subunits. However, in general, folds are not correctly predicted even for average size monomeric targets. This is the case even though the community as a whole makes about 500 predictions. Perhaps new conceptual or algorithmic leaps will be required to achieve greater coverage of targets that meet some "near-correct" standard.

Evaluation of assessment criteria

We took advantage of the CASP experiment to assess not just the models, but also the criteria that are used to assess the models. For 13 of the 18 targets, at least one model was independently selected by all three assessors in the second round of assessments. In cases where there was no unanimity, it was often because the differences among the best models were subtle. In other cases, individually chosen models were substantially different. Such discrepancies highlight the fact that assessors had different subjective priorities about what constitutes a good model. This was especially the case when overall model-target similarity was poor. When averaged over all 18 targets, GDT_TS yielded a ranking of prediction groups that is quite consistent with what was seen by visual assessment. This suggests that GDT_TS might be sufficient if the only goal is to identify the best groups. However, GDT_TS is not sufficient if it is desirable to identify the best models for each target.

Typically, the best models by visual inspection ranked high by GDT_TS, but in many cases the best GDT_TS model was not among the best visual models. This was not a case of the best GDT_TS models somehow slipping through the cracks because we showed, in a blind reassessment, that assessors nearly always (14/15 times) picked again the top-scoring visual model rather than the top-scoring GDT_TS model. In general, the preference for the best visual models reflects their (subjectively) more accurate topology while the models identified by GDT_TS may be biased toward models with more accurate local structure. In a couple of cases, the subjective preference for the best visual model is readily rationalized. For example, for T0321_D2, only the best visual model had the right orientation of a particular β-strand and for T0307, the best GDT_TS model had a superhelical arrangement of helices that was clearly different from both the target and from the model judged best by visual inspection.

An important question for future assessments is how far down the GDT_TS list one needs to go to be reasonably confident that a good model by visual criteria will not be missed. In CASP7, about half of the models favored by two or more assessors were found among the top 10 GDT_TS models (see Fig. 11). More than 90% of the models identified by two or more assessors can be found in the top 25 GDT_TS models, and essentially all are found in the top 50 (see Fig. 11). We also used a contact map similarity measure but found it to be less well correlated with visual assessments than was GDT_TS. However, contact map similarity did have some value as a method complementary to GDT_TS for prioritizing models for visual inspection. Perhaps other objective measures of target-model similarity can be developed in the future that will prove better correlated with subjective assessments. Until then, it appears that visual inspection will have to be a part of FM assessment.

ACKNOWLEDGMENTS

We thank the organizers for the invitation to contribute to CASP, and to the CASP Prediction Center, especially Andriy Kryshtafovych, for the fantastic infrastructure support and advice that made this assessment possible.

REFERENCES

1. 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.
2. 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.
3. Lesk AM, Lo Conte L, Hubbard TJ. Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, and interresidue contacts. Proteins 2001;45(Suppl 5):98–118.
4. Jones DT. Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. Proteins 1997;29(Suppl 1):185–191.
5. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 1998;281:565–577.
6. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
7. Grana O, Baker D, MacCallum RM, Meiler J, Punta M, Rost B, Tress ML, Valencia A. CASP6 assessment of contact prediction. Proteins 2005;61(Suppl 7):214–224.
8. Delano WL. The PyMOL molecular graphics system. 2002.