Assessment of CASP6 Predictions for New and Nearly New Fold Targets

James J. Vincent, Chin-Hsien Tai, B.K. Sathyanarayana, and Byungkook Lee*
Laboratory of Molecular Biology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland

ABSTRACT **This is a report of the assessment of the predictions made for the CASP6 protein structure prediction experiment conducted in 2004 in the New Fold (NF) category. There were nine protein domains that were judged to have new folds (NF) and 16 for which a similar structure was known but the sequence similarity was judged to be too low for them to be easily recognized (FR/A). We selected all NF targets and eight of the 16 FR/A targets judged to be at the borderline between NF and FR/A for evaluation in the NF category. A total of 165 prediction groups submitted over 7400 structural models for these targets. The quality of these models was evaluated using the GDT_TS scores of the structural similarity detection program LGA and by visual inspection of the top-scoring models. The best models submitted bore an overall similarity to the target structure for three or four of the nine NF targets and for all but one of the FR/A targets. High-scoring models for the NF targets were submitted by several different groups. When both the NF and FR/A targets were considered, Baker group dominated by submitting best models for seven of the 17 targets, but 14 other groups also managed to submit best models for one or more targets. Proteins 2005;Suppl 7:67–83. © 2005 Wiley-Liss, Inc.***

Key words: CASP6; protein structure prediction; New Fold prediction assessment

INTRODUCTION

The Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiment has been enormously useful, both for protein structure predictors and for the scientific community in general, because it provides an objective assessment of the effectiveness of structure prediction methods and a sense of the state of the art in protein structure prediction. Predictions are made on proteins for which the structure is unknown but the structure determination is imminent. The predictions are collected over a half-year period and the quality of the predicted models is assessed against the newly determined structures. The experiment is conducted every other year and in an open fashion so that anyone can participate in the experiment.

This is the report of the assessment of the predictions made for the CASP6[1] experiments conducted in 2004 for the targets that are in the New Fold (NF) and difficult Fold Recognition Analogous (FR/A) categories. (See Dunbrack et al.[2] for the fold class assignment.) The assessment of the NF category is to evaluate the effectiveness of the techniques used for the case when a similar structure does not exist in the database. However, many FR/A targets are successfully predicted using techniques that are effective in predicting the NF targets. For this reason, we initially evaluated predictions of all targets within both NF and FR/A categories. Since predictors did well for targets for which a closely similar target exists, we also carried out the analysis after excluding these "easy" FR/A targets. Here we report only on the results of evaluating the nine NF and the eight "difficult" FR/A targets. The latter are those that have the average GDT_TS score (see below) at or below 25.

We evaluated the quality of predicted models using a quantitative measure of similarity and by visual inspection. For a quantitative measure, we used only the GDT_TS scores calculated by the structure comparison program LGA[3] and provided by the CASP6 prediction center.[1] The GDT_TS scores were used as a primary measure of model quality in CASP5 NF assessment.[4] We found that GDT_TS scores reflected the quality of models, as judged by visual inspection, quite well in most cases. Exceptional cases are noted in this paper.

For visual assessment, the models were sorted in descending order of the GDT_TS score for each target and inspected visually (by BL) from the top of the list. A large number of models were scanned quickly, but usually only a relatively small number of models at the top could be examined in detail. The actual number studied in detail varied by target. We noted only the top quality model(s) for each target from this inspection.

We devote most of the paper to the description of the visual inspection of models for individual targets and relatively little for analysis using the GDT_TS scores, since the latter are available at the CASP6 web site[1] for an analysis by any interested person. We must also caution

Grant sponsor: NIH, National Cancer Institute, Center for Cancer Research

*Correspondence to: B.K. Lee, Bldg. 37, Rm. 5120, National Institutes of Health, 37 Convent Drive MSC 4264, Bethesda MD 20892-4264. E-mail: bk@nih.gov

Received 28 May 2005; Accepted 20 June 2005

Published online 00 Month 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20722

This article was originally published onine as an accepted preprint. The "Published Online" date corresponds to the preprint version.

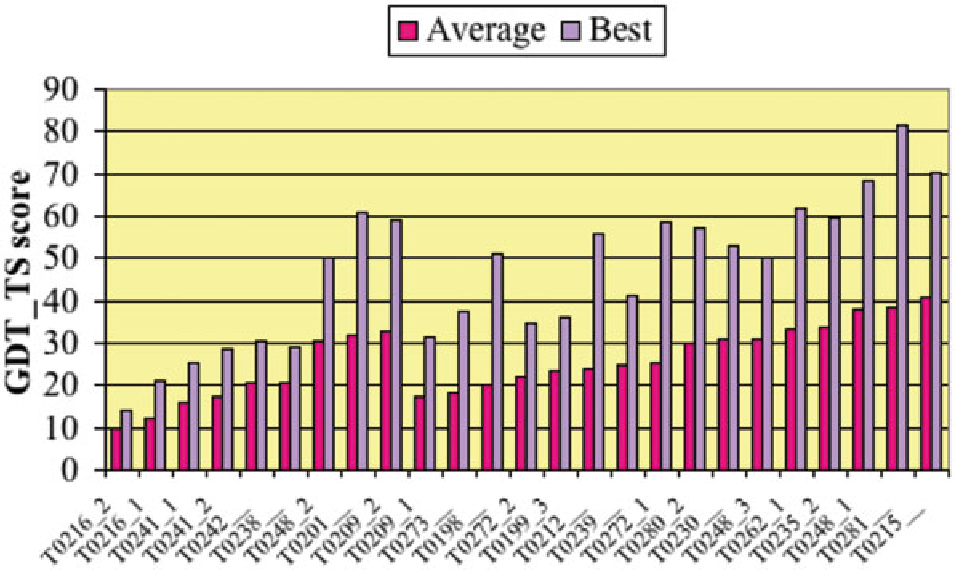

Fig. 1. Average and best GDT_TS scores for all NF and FR/A targets.

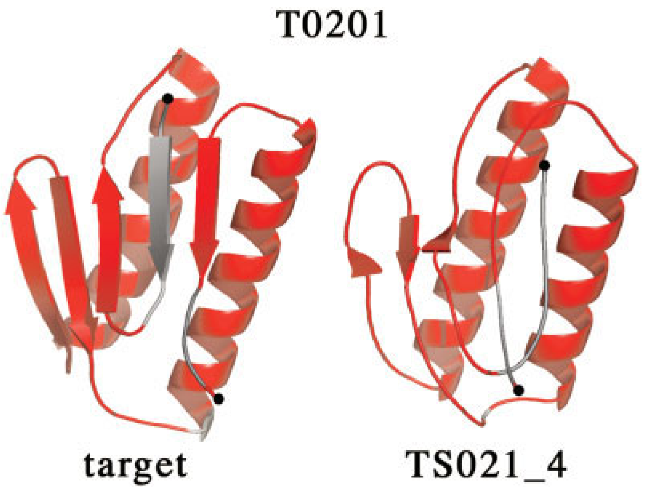

Fig. 2. Target T0201 and model TS021_4. Aligned regions between model and target are colored red, unaligned regions are colored grey. Secondary structure was assigned by DSSP program. Alignment is based on 6-Å cutoff from LGA alignment distances.

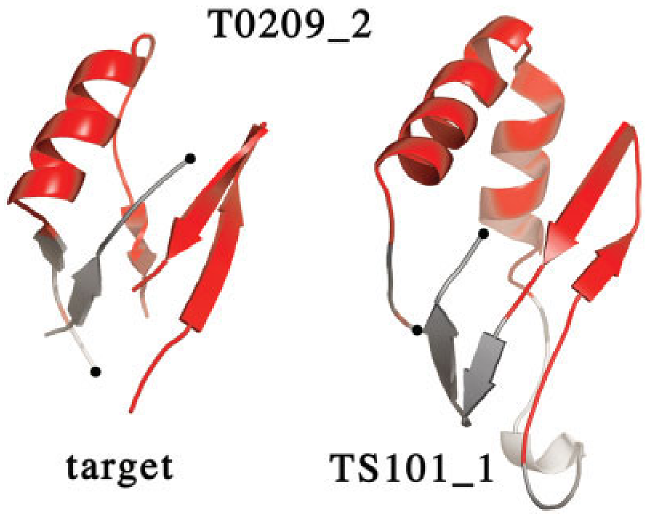

Fig. 3. Target T0209 domain 2 aligned to model TS101_1. The alignment, the secondary structure assignment, and the color scheme, are as described in Figure 2.

that the sample size is rather small—there are only 17 targets total—for drawing statistically robust conclusions from numerical scores alone.

MATERIALS AND METHODS

Visual Inspection

We used pyMol[5] for a rapid scan through a large number of predicted structures. For a detailed visual examination, we used an in-house molecular graphics program GEMM. This program has a facility to superimpose two structures using an input alignment, or manually by using dials. The superimposed structures were viewed in stereo. For the alignment between a model and the target, we mainly used the LGA alignment supplied by the CASP Prediction Center, but in many cases we also sought alternate alignments by superimposing them manually.

Templates were examined for some targets in order to understand the errors in the prediction. These were supplied by Drs. Tress and Valencia using LGA and by searching through the ASTRAL database[6,7] using another structure–structure alignment program SHEBA.[8]

GDT_TS Score Analysis

Models used

Predicting groups submitted up to five models for a given target. We selected only two: the model labeled number one by the submitting group (model one) and the model that had the highest GDT_TS score (max model). All

TABLE I. The Average and Best Scores and Visually Best Models for Each Target

Target ID[a]	Size (AA)	Class[b]	Average[c] GDT_TS	Best[c] GDT_TS	Best models by visual inspection[d]
T0216_2	213	NFh	9.9	14.0	100_2*
T0216_1	213	NFh	12.0	21.3	021_1
T0241_1	117	NFh	16.2	25.4	060_1, 003_3*, 604_4*
T0241_2	119	NFh	17.3	28.8	113_1
T0242	115	NFh	20.5	30.7	003_5
T0238	153	NFh	20.8	29.3	172_1*
T0248_2	87	NFe	30.6	50.0	101_2*, 450_2*
T0201	90	NFe	32.1	61.2	021_4
T0209_2	73	NFe	33.5	59.2	101_1, 501_1, 021_3, 450_2
T0209_1	130	FR/A	17.4	31.5	185_2
T0273	186	FR/A	18.1	37.5	100_2
T0198	221	FR/A	20.2	51.1	100_2
T0272_2	122	FR/A	21.9	34.6	100_2
T0199_3	82	FR/A	23.7	36.3	089_4
T0212	119	FR/A	23.8	55.8	100_2
T0239	98	FR/A	25.0	41.3	530_2, 035_2*, 157_1*
T0272_1	85	FR/A	25.5	58.5	100_1

[a] CASP6 target codes. Underscore appended numbers are the domain numbers.

[b] CASP fold class. NFe and NFh indicate "easy" and "hard" NF targets.

[c] Scores are for the max model set.

[d] Predicted models are indicated by the predictor ID number followed by an underscore followed by the model number. For the group number–name correspondence, see Table II. *Not the top GDT_TS scorer.

other submissions from a group for each target were ignored for the quantitative analysis. All rankings and analyses were made separately for the set of max models and for the set of model ones.

Average score across different targets

In order to assign a performance measure to each predicting group, we computed the group's average GDT_TS score across different targets. Targets with no submission were assigned a score of 0. The average was then calculated over all targets regardless of the number of actual submissions the group made. The Z-score (see below) average was computed similarly.

Average Z-scores

The Z-score, $Z(g, t)$, of a model from group g for target t was calculated as

$$Z(g,t) = \frac{s(g,t) - <s>_t}{\sigma_t},$$

where $s(g,t)$ is the raw GDT_TS score of group g for target t, $\langle s \rangle_t$ is $s(g,t)$ averaged over all models submitted for t, and σ_t is the standard deviation of $s(g,t)$ about $\langle s \rangle_t$. Negative Z-scores were set to zero before computing the average Z-score for each predicting group.

RESULTS

Targets Selected for Evaluation

The average and best GDT_TS scores for each target from the max model set (see Methods section) are shown in Figure 1 for all NF and FR/A targets. Targets included in the analysis reported here are the nine NF targets and the eight FR/A targets with average GDT_TS scores at or below 25. This number was chosen primarily because there was a relatively large gap in the score between T0272_1 (GDT_TS = 25.5) and T0280_2 (GDT_TS = 30.2). The average and best GDT_TS scores for each target from the max model set are given in Table I, along with the target size and class.

Table I and Figure 1 show that NF targets T0248_2, T0201 and T0209_2 have average GDT_TS scores above 30. We refer to this group of three as NFe (e for "easy") targets and the other six as NFh (h for "hard"). We thus have three target sets: six NFh, three NFe, and eight FR/A. Quantitative scores (GDT_TS or Z-score) were summed or averaged separately for the three sets and for all targets.

Visual Inspections for Individual Targets

Visual inspection results are described below for each target. With each target ID are given the PDB file name, the residue number range, the prediction class, and SCOP-style structural class designations.[9] In describing the structure of a target or model, secondary structural elements (SSEs) are indicated by a Greek letter followed by a serial number. The Greek letter α represents both alpha and 3_{10} helices, and the letter β represents strands. The number always refers to the order in which the SSE occurs in the sequence. When two numbers are given with a hyphen between, they represent a range of residues.

T0198 [1SUM (3–223), FR/A, α]

This is an up-and-down six-helix bundle. Both of the N- and C- terminal halves belong to the pfam[10,11] PhoU domain. There is a short β-hairpin at the C-terminal end. There is no suitable template that has the six-helix bundle structure of this target, but there are many proteins that have the up-and-down three-helix bundle motif, from which the six-helix bundle structure can probably be built without resorting to a template-independent folding technique.

The top-scoring model is from the Baker group (TS100_2). This is an excellent model with all the helices at nearly the correct position and orientation with respect to each other. A noticeable imperfection is that the two three-helix sheets are shifted with respect to one another by about one helix width. Also, the C-terminal tail is modeled as a short helix hairpin rather than as a short β-hairpin. The second high scoring model is also from the Baker group (TS100_1). It is nearly as good as the top-scoring model, but the helices are bent a little more than in the target. These two models have substantially higher GDT_TS scores than all other models. Indeed, all other top scoring models examined have the helices arranged in an obviously incorrect manner and the C-terminal β-hairpin is not modeled any better.

T0199_3 [1STZ (145–226), FR/A, α + β]

The first 35 residues (145–179) form a three-stranded β-sheet of simple up-and-down topology in this structure and the rest forms a structure that consists of two short and two long helices. There are many similar structures, e.g., d1muca2, in the SCOP[9,12] d.54.1.1 family.

Surprisingly, no one found the correct template. The best scoring model is TS089_4 from KIAS group. This model matches only the three strands of β-sheet very well. The helices α1 and a combination of α2 and α3 do exist in the model but are quite misplaced. A number of other models also have only the β-sheet part correct. The second best scoring model is TS305_2 from the Ho-Kai-Ming group. This model and several other models have only the β2, β3 and α1 nearly correctly modeled. The tenth ranking model, TS009_4 from UGA/IBM-PROSPECT group, includes a good model of α2, α3, and α4, while the fourteenth ranking TS031_3_2 from PROTINFO-AB is a pretty good model of the whole α-helical part of the domain. Thus, different high scoring groups managed to obtain some parts of the model correctly, but no group could assemble or recognize the correct overall structure for this domain.

T0201 [1S12 (1–90), NFe, α + β]

This structure is an $\alpha_2\beta_5$ two-layer sandwich (Fig. 2). No good template was found but the target structure can be inserted into the structure of d1kija2 in the SCOP superfamily d.122.1.2 to obtain a good match for all five strands and a fair match for one of the two helices.

The top-scoring model from Kolinski-Bujnicki's group (TS021_4, Fig. 2) is excellent. The whole model of residues 1–94 can be superimposed to the target structure with Cα root-mean-square deviation (RMSD) of 3.5 Å. But the first and the last β-strands are swapped in space. The second high-scoring model (GDT score 51.06) is model one from the same group. This is also a good model with all the strands and helices topologically correctly arranged, except for the first and last strands, which are again swapped. This model does not fit the target as well as their model 4 primarily because the angle between the strands and the helices is less accurate. Many other high-scoring models are good in that they are made of a two-helix layer and a β-sheet layer, that the helices are anti-parallel, and that the β-sheet is on the correct side of the α-helical layer. However, many have a wrong number of strands and/or an incorrect topology of the β-strands in the sheet. A couple of models from the Baker group (TS100_3 and TS100_1) and a model from the SAM-T04-hand group (TS166_1) do have the correct β-strand connectivity. But their quality is not as good as the top-scoring model; some helices are distorted, the β-sheet is distorted, and/or the relative position and orientation of the helices with respect to each other and with the β-sheet are inaccurate.

T0209_1 [1XQB (9–138), FR/A, β]

This is a six-stranded Greek-key β-barrel. Residues 64–85 in the linker between β3 and β4 are missing in the target structure. Many known structures have this same basic fold. Examples found by SHEBA are d1kk1a1, d1gvha2 and others in the SCOP b.43.3 and b.43.4 superfamilies.

The predictors found this domain to be the most difficult among FR/A targets; the average GDT_TS score for this domain is lower than half of the NFh targets. Rather surprisingly, only three groups (UGA/IBM-Prospect's TS009_3, Eidogen-EXPM's TS223_1, and CBRC-3D's TS272_5_1) indicate that they used a template in the b.43.3 or b.43.4 SCOP superfamilies. All are poor models with low GDT_TS scores.

The top-scoring model is TS185_2 from Huber-Torda's group. This model was derived from d1ejea_, which belongs to the SCOP b.45, the "FMN-binding split barrel" fold. The model is quite similar to the template used and the major differences between the model and the target arise directly from the differences in the two folds, b.45 and b.43. These include alignment shift of some 20 residues for β1 and wrong placement of the C-terminal residues corresponding to β6.

Other models are noticeably poorer, which we do not describe here since this is primarily an FR target, except for one feature. Model TS101_1 has an excellent structure for domain2, but a poor one for domain1 of this target; the GDT_TS score for domain1 is 23.15, at rank 10. It is mentioned in discussing T0209_2 that the groups P021 and P450 appear to have adopted this model for domain2, but not for domain1. The former managed to replace it with a better model (TS021_3, GDT_TS 27.08, rank 3), the latter with a poorer one (TS450_2, GDT_TS 22.22, rank 14).

T0209_2 [1XQB (159–231), NFe, α + β]

We do not know the true structure of this domain since two pieces in this domain, between α2 and β1 and between β3 and β4, and pieces from the N- and C-termini are missing. The visible parts of the structure are shown in Figure 3. There is no good template for the whole structure, but there are structures in the SCOP fold d.110.4.2 superfamily, e.g., d1gw5s_ and d1h8ma_, that contain substructures that match a large part of the target structure.

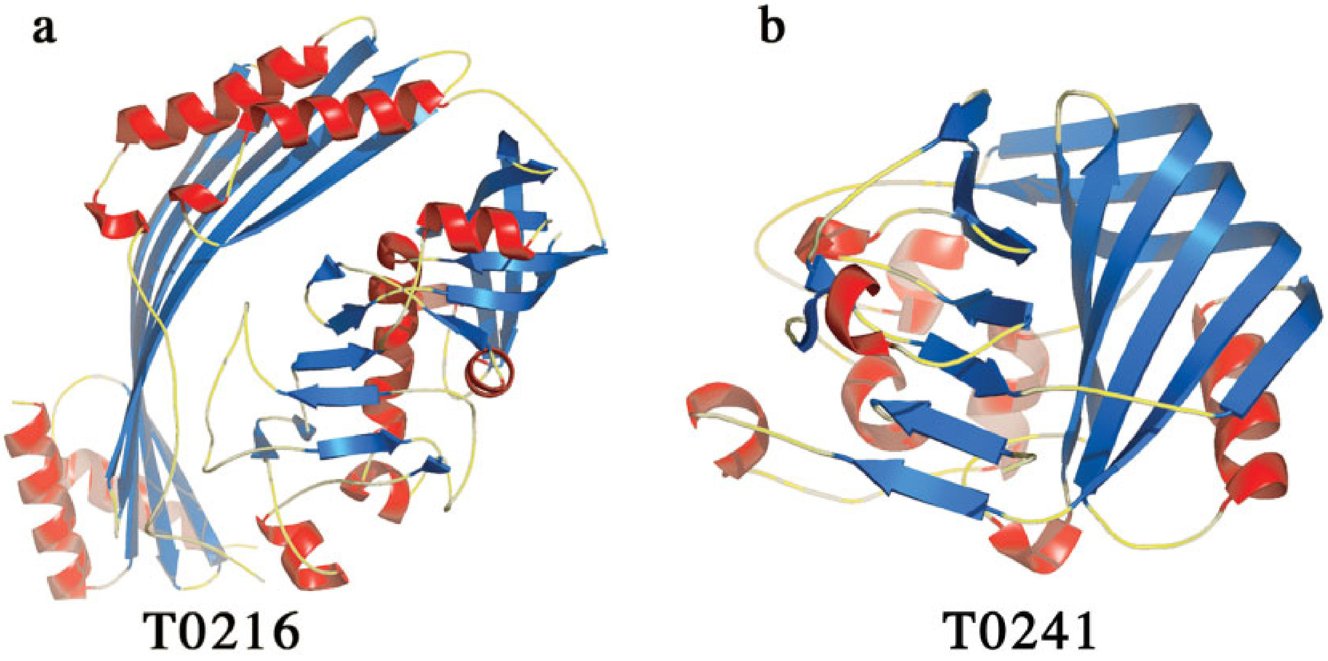

Fig. 4. Two most difficult NF targets proteins: (**a**) Target T0216, (**b**) Target T0241.

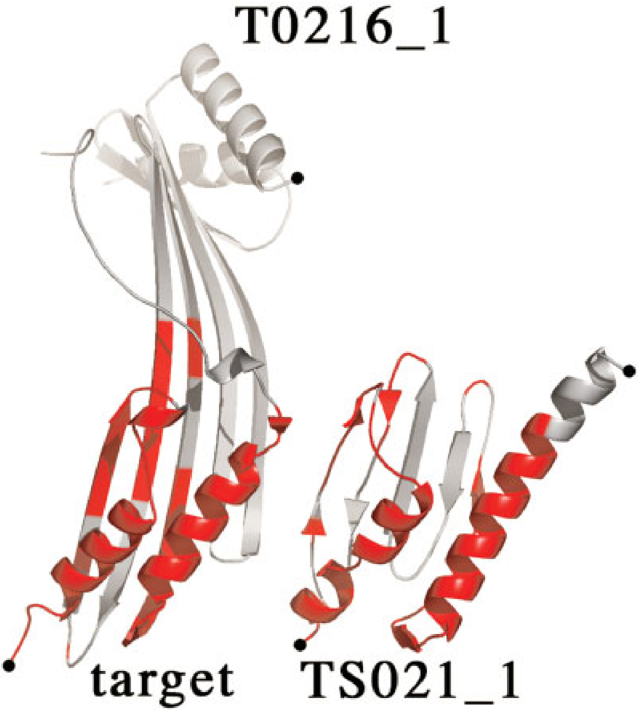

Fig. 5. Target T0216 domain 1 aligned to model TS021_1. The alignment, the secondary structure assignment, and the color scheme, are as described in Figure 2.

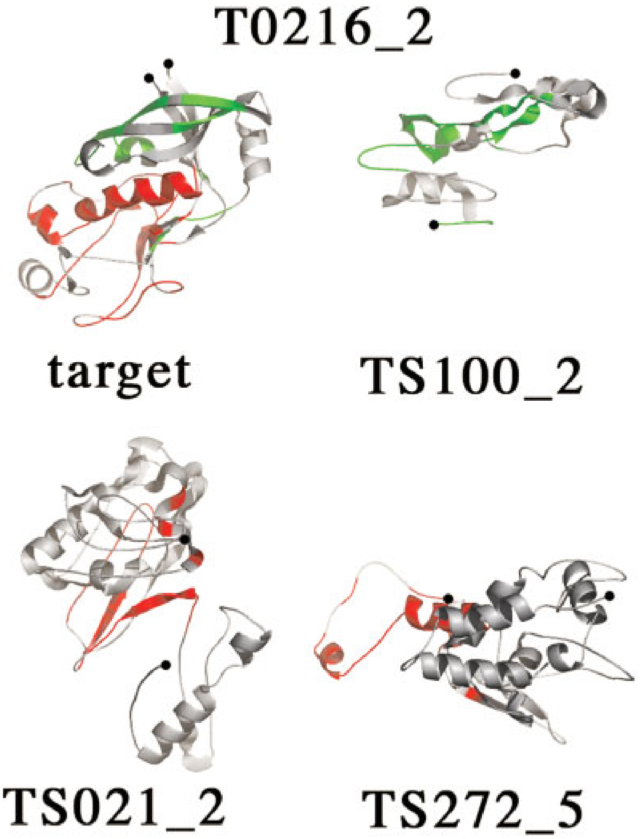

Fig. 6. Target T0216 domain 2 aligned to models TS100_2, TS021_2 and TS272_5. Aligned regions between target and model TS100_2 are colored green. Aligned regions between target and models TS021_2 and TS272_5 are colored red. Unaligned regions are colored grey. The secondary structure assignment and the alignment are as described in Figure 2.

The top-scoring model is an excellent model. The entire visible domain, which consists of residues (159–192), (206–220), and (224–231), can be superimposed to the target with an RMSD of 4.4 Å between Cα atoms. Interestingly, four top-scoring models were identical. Model TS101_1 is from Baker-Robetta server and model TS501_1 from Mcon group is identical to it for both domains of this protein. The Mcon group is experimenting with a method to select the best from a set of models predicted by other servers.[13] The method was clearly successful in this case. The other two models are TS021_3 (Kolinski&Bujnicki) and TS450_2 (Ginalski). These are identical only for domain 2; each has its own, unique domain 1.

Of the top 25 models, 12 are from Baker's laboratory (Baker, Baker-Robetta, and Baker-Robetta_04) and four others are identical to some of these. The remaining nine were submitted by just four different groups, SAM-T04-hand, Keasar, Ginalski, and Rokko. These are all good models, with the two helices and four-stranded β-sheet placed and oriented nearly correctly. The quality of the models decreases as the GDT_TS score decreases.

T0212 [1TZA (3–121), FR/A, β]

This is a seven-stranded β-sandwich of the immunoglobulin fold. Similar structures can be found, for example, in SCOP b.1.5 superfamily.

The top-scoring model is TS100_2 from the Baker group. This is an excellent model wherein all the important features of the target structure are quite accurately placed, except β1, which is misplaced. Even the β-bulge in β4 is reproduced, although misplaced by two residues. The top models from CBRC-3D (TS272_1) and from Jones-UCL (TS003_1) are both also good models wherein all seven β-strands are essentially correctly placed.

T0216_1 [1VL4 (2–214), NFh, α + β]

T0216 is a 435-residue, officially two-domain protein of novel fold [Fig. 4(a)]. The N-terminal domain is nearly symmetrical; the N-terminal half of the domain is related to the C-terminal half by an approximate two-fold symmetry. A long linker, a part of which is missing in the crystal structure, connects the two halves. Each half is made of a simple up-and-down three-stranded β-sheet with two α-helices on one side of the sheet, in the αβββα sequence. Remarkably, the first two β-strands are each about three times as long as the third strand or the two helices and form a continuous β-sheet with the corresponding parts from the other half. Thus, the top and bottom thirds of the whole domain structure are each made of a five-stranded sheet plus two helices while the middle third is made of a four-stranded sheet.

The basic αβββα structure can be found in the SCOP d.50.2.1 family of proteins, e.g., d1pda_2. However, they do not have the extra long two β-strands or the large β-sheet, which is the most striking feature of this structure.

Many top scoring models contain the β-sheet-and-two-helices motif at least for one subdomain. However, in most models, the long β-hairpin is folded in half and tucked in between the stems of the first and second strands of the target structure, producing a five-stranded sheet. Exceptions are models 4 and 5 from the Baker-Robetta server (TS101_4 and 5), which do contain the long hairpin in one of the two halves. The five-stranded model actually looks superficially similar to the top or the bottom end of the target structure, as illustrated in Figure 5 by the top-scoring model from Kolinski&Bujinicki group (TS021_1), but obviously the topology of the connection is not similar. This model also has the two halves of the target domain as independent domains with no interaction between them. However, the two domains have a similar structure as in the target structure. In each domain, the β-sheet and the angle between the strands and the helices are also modeled rather accurately. The model shown in the figure is the C-terminal half. One other model, TS035_1 from the GeneSilico group, also has basically symmetric two halves, which in addition interact with one another by forming a continuous β-sheet. But the relation between the two halves is translational (β1 of first half interacting with β5 of the second) rather than rotational (β1 interacting with β1).

T0216_2 [1VL4 (221-433), NFh, α + β]

The predictors found the structure of this 213 amino acid domain to be the hardest to predict. This domain is made of two subdomains (Fig. 6). The core structure of the subdomain D21 (218–235 and 338–433) is a β-barrel, which includes one β-strand from the amino terminus of the domain but the rest of which is made of the carboxy terminal residues. The core of the other subdomain (236–337) is a three- or four-stranded β-sheet. But the domain contains a large number of other residues, which form loops and two small helices outside of the core β-sheet. The two subdomains interact intimately through a long β-hairpin loop (408–428) that hydrogen-bonds with, and thus extends, both the β-barrel of D21 and the β-sheet of D22. There is also a prominent helix (α1, a mixture of α- and 3_{10} helices, 228–243) that spans both domains. There is what appears to be a useable template for D21: the residues 30–90 of 1onc, and corresponding residues in other ribonucleases, mimic residues 349–399 of D21 reasonably well, although there is no sequence homology.

The facts that the domain is a part of a large protein (433 amino acids), that the domain is made of subdomains that interact intimately, and that one of the subdomains is made of residues at opposite ends of the domain probably contribute to the difficulty of correctly predicting this structure. The best GDT_TS score for this target was $< 14\%$. We carefully inspected more than a dozen submitted models that had top GDT_TS scores for this domain or for the entire T0216 structure, but could not find any model even approximately similar to the target structure. The four top-scoring models (TS272_5, TS052_2, TS052_5, TS176_3) attained the high GDT_TS score almost entirely by virtue of having the α1 helix (Fig. 6). The fifth ranking model (TS021_2) has some loops that resemble the target. The sixth ranking model from the Baker group (TS100_2) has a more convincing fragment (365–398) that includes a short helix and a couple of β-strands that form a part of the barrel of D21. This fragment is within the region mimicked by 1onc, but the model does not resemble this structure.

T0238 [1W33 (70–222), NFh, α]

This structure (Fig. 7) is basically made of an up-and-down bundle of three helices (α1, α5, and α6), connected in

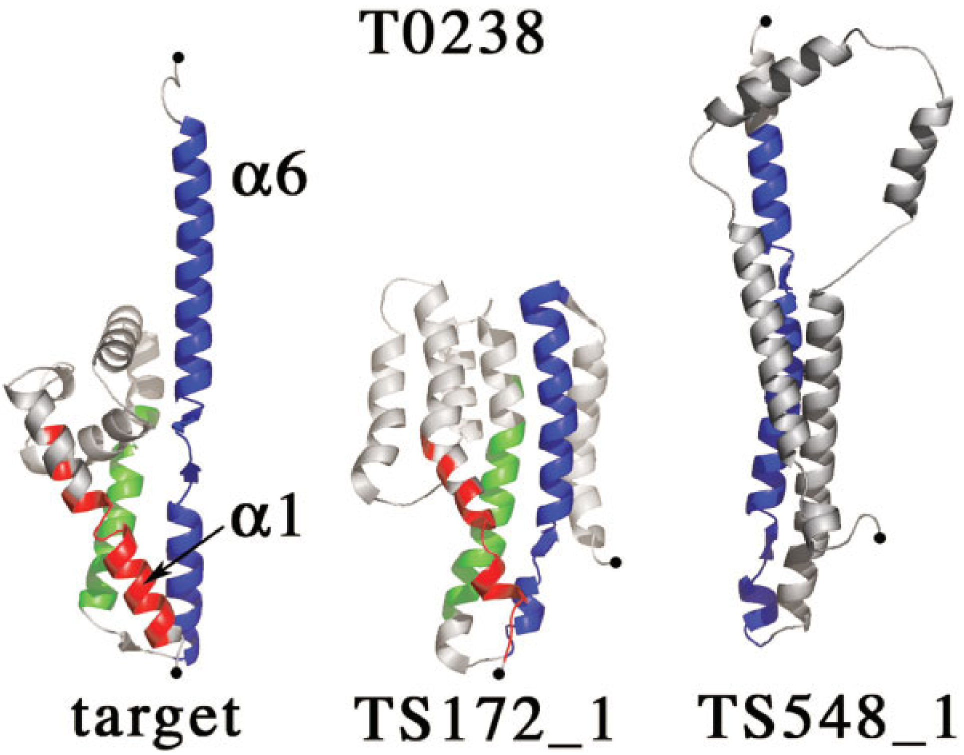

Fig. 7. Target T0238 aligned to models TS172_1 and TS548_1. Aligned regions between target helix α1 and models are colored red, between target helix α6 and models are colored blue, and between target helix α5 and models are colored green. Unaligned regions are colored grey. The secondary structure assignment and alignment are as described in Figure 2.

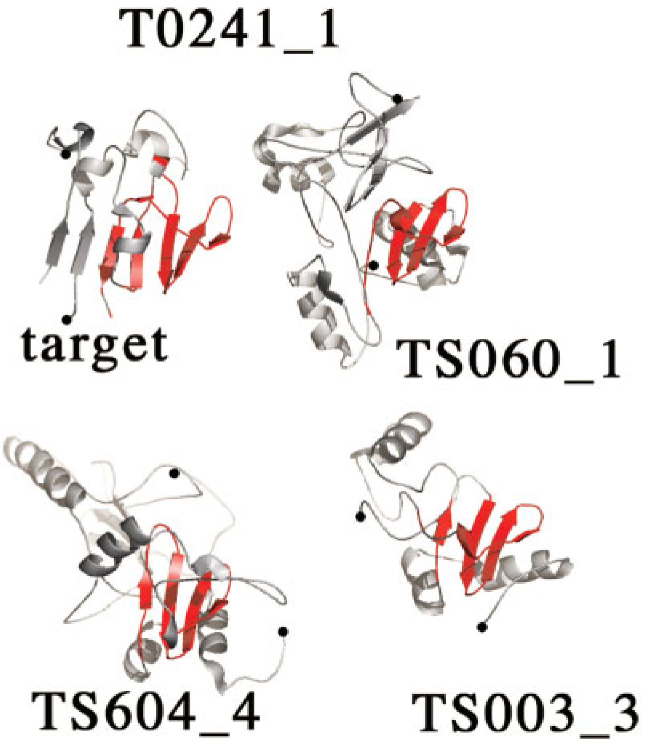

Fig. 8. Target T0241 domain 1 aligned to models TS060_1, TS604_4, and TS003_3. The alignment, the secondary structure assignment, and the color scheme, are as described in Figure 2.

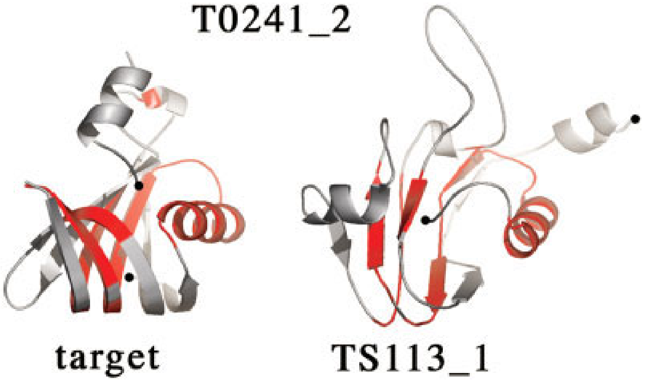

Fig. 9. Target T0241 domain 2 aligned to model TS113_1. The alignment, the secondary structure assignment, and the color scheme, are as described in Figure 2.

the left-handed manner. But one end of the bundle is opened up to accommodate a set of two well formed (α3 and α4) and one ill-formed (α2) shorter helices between α1 and α5. These shorter helices are nearly perpendicular to the main helices. Nearly half of α6 juts out beyond the compact region.

The top-scoring model from the JIVE group (TS548_1) attained the high GDT_TS score (29.28) only by having the long α6. Other parts of the structure do not resemble the target. The three major helices do form a helix bundle structure, but of the right-handed topology. The α6 is modeled well, however, since its GDT_TS score is considerably better than that of, for example, TS229_3 (GDT_TS 25.27, rank 23), which is made entirely of one long helix.

Because it is possible for this target that a better model can have a lower GDT score if its α6 is less perfect, we visually inspected more than 50 top scoring models for this target. As expected, many of these achieve the relatively high GDT score because they have the long α6, with little else correct. Others had a less correct α6, but were helped by correctly arranged α5. The best model was clearly TS172_1 from the ProteinShop group (Fig. 7), which had the second best GDT score (29.14). This model has all three main helices approximately correctly arranged, although α6 is broken in half and its C-terminal end is folded back along the side of its N-terminal part. However, the three-helix bundle is not opened up at one end and the middle three shorter helices are modeled incorrectly.

T0239 [1RKI (1–98), FR/A, α + β]

This is an $\alpha_2\beta_4\beta_3$ three-layer structure. The β_3 layer is at the C-terminus of the protein and forms a small separate domain (D2, 71–98). The $\alpha_2\beta_4$ main domain (D1) has a ferredoxin-like fold, with the SSEs arranged in βαββααβ sequence.

A popular template for this structure is d1xxaa_. However, this and other similar structures resemble a circularly permuted form of D1, wherein the C-terminal residues corresponding to β4 (65–70) are cut and placed at the N-terminus. None of the models we examined has a good structure beyond residue 64. Many models do not contain residues beyond 64. This suggests that these template structures had a strong influence in building models for this target. For example, 10 of the 15 best scoring models, including the top-scoring TS530_2 from the B213-207 group (GDT_TS = 41.33), have a good alignment only for residues up to 47, the last residue of β3, even though all have the helical segment following it at the position of α2 and α3. This is because these models have five extra residues in the loop between β3 and α2, a feature that exists also in the d1xxaa_ template, and the alignment is off by the same amount for all subsequent residues. Four other models have the correct length for the β3–α2 loop, even though d1xxaa_ was used as a parent for at least two of these models, and the alignment is extended to include α2 and part or all of α3. The best models among these are probably TS035_2 from the GeneSilico group (GDT_TS = 37.76, rank 3) and TS157_1 from the 3D-JIGSAW group (GDT_TS = 36.22, rank 14), which are of comparable quality.

For D2, we could spot only one model that had a basically correct partial structure: AL164_3 from SAM-T02 group, cut from 1dfeA. This model is nearly at the bottom in rank because it has only 19 residues.

T0241_1 [1XV2(1–79,127–144,205–224), NFh, α/β]

T0241 is a 237-residue, officially two-domain protein of novel fold [Fig. 4(b)]. This first domain is made of a seven-stranded twisted β-sheet with one short helix on one face, and three others on one edge, of the sheet (Fig. 8). The domain is made of three segments: the main segment (1–79) is from the N-terminus, one of the edge helices is from the middle of the sequence (127–144) and one strand and another edge helix are from near the C-terminus (205–224) of the whole target protein. Counting from the C-terminal strand, the order of the strands in the sheet is 7-1-6-2-3-4-5. The last four strands are connected by hairpin loops between neighboring strands and arranged in a simple up-and-down manner.

The predictors found this to be a difficult target: no model examined was globally similar to the target structure. All but one of the top-scoring 10 models contained a nearly correct substructure that roughly corresponds to the simple up-and-down portion of the β-sheet. For example, the top-scoring model from the Bilab group (TS060_1) has a quite accurate portion, which corresponds to the β-strands β3-β4-β5 (residues 44–68). The residues 69–79 are also approximately correctly modeled, except that this strand (β6) is placed next to β3, whereas it is between β1 and β2 in the target structure. The second and third high scoring models are from Baker (TS100_5) and Baker-Robetta_04 (TS604_4) groups and similar. These and another from Jones-UCL (TS003_3) group are quite similar in the parts that match the target structure. They have nearly the same GDT_TS scores as the top-scoring model (25.4, 25.0, 24.8, and 24.2) and nearly the same or more (TS003_3) number of matching residues, which include β2-β3-β4-β5 (residues 33–67 for TS100_5 and TS604_4 and 33-75 for TS003_3). In all four models, the N-terminal residues between β1 and β2 are placed on the wrong side of the sheet.

T0241_2 [1XV2 (81–226,145–204,225–237), NFh, α/β]

The core of this domain is a five-stranded β-barrel, which is closed at one end and open at the other (Fig. 9). The open end is covered by a helix (α1). The domain is made of three segments: the N-terminal residues (81–126) that form two strands of the barrel and α1, the middle residues (145–204) that form the rest of the barrel structure, and the C-terminal residues (225–237) that form an extra helix that does not directly interact with the core structure, but interacts with a loop that protrudes out of the open part of the barrel core.

SHEBA identified several structures in the SCOP b.121.4 superfamily as having a substantial number of structurally matching residues. The structure, d1b35b_, for example, matches the four strands in the open part of the barrel and the α1 helix rather well. It is, however, a β-sandwich, jelly roll type of structure and does not form a closed barrel.

This is another difficult target for which no model was submitted that had the correct overall shape. The three top-scoring models (TS113_1, TS109_2, TS506_2) are all from the SBC group and similar to each other. The two automatic submissions (TS113_1 and TS506_2) used 1qqp as the parent, which also belongs to the SCOP b.121.4 superfamily. In fact the models are quite similar to the parent and share the similarities and dissimilarities of d1b35b_ described above. Other models are considerably worse than the top three models. In particular, none has the closed barrel, which is the part not modeled by the top three.

T0242 [2BLK (2–116), NFh, α/β]

This is basically a five-stranded β-sheet, with two nonparallel helices on one side, a couple of irregular loops on the other side, and a short helix nearly in the plane of the sheet (Fig. 10). The strands in the sheet are in 2-5-4-1-3 order and all linkers between them are long. The long linker between β1 and β2 is one of the irregular loops and includes one 3_{10} turn. Between β2 and β3 is the main helix (α1), which is at about 45° angle to the strands and spans the whole sheet. The linker between β3 and β4 initially forms the second helix (α2) on one side of the sheet, but then winds around to the other side of the sheet to form the second irregular loop. The linker between β4 and β5, the only pair that are sequentially adjacent and antiparallel to each other, is flared out and includes a short helix (α3), nearly in the plane of the sheet. No good template with the correct topology could be found.

This was also a difficult target for the predictors. The best scoring model was from Jones's UCL group (TS003_5, GDT_TS = 30.65). This model has an approximately correct shape and matches β1, β4 and α3 well. The β2 and β5 are also nearly correctly modeled, although they are somewhat misplaced. On the other hand, α1 is totally missed and the residues 72–77, which form the linker between α2 and β4 in the back of the sheet in the target structure, are placed where β3 is in the reverse orientation (Fig. 10).

Other high scoring models are considerably worse but some have remarkable features. One of the second best scoring models (TS237_2, GDT_TS = 28.26) has a structure that looks very different from the target (Fig. 10). The high GDT score was obtained because (1) the model has three helices at the correct position in the sequence and (2) these are placed approximately at correct relative positions in the model. However, the connecting structures between the helical elements are different from that in the target. This is a rare example that shows that three SSEs can be placed nearly correctly even when the whole structure is different and that the GDT score sometimes does not reflect the true quality of the model.

Another model from the same group (TS237_4, GDT_TS = 26.09) is considerably better although it has a lower GDT score. Residues 93–111 are modeled very accurately—RMSD is only 0.88 Å. Remarkably, these are the flared-out linker residues between β4 and β5. Residues 28–61, which span α1 to α2, are also modeled correctly although much less accurately than (93–111). But the relative orientation of these two fragments is wrong so that one cannot align both fragments together.

T0248_2 [1TD6 (107–193), NFe, α + β]

This is an up-and-down bundle of four helices and a β-hairpin loop. There are many structures, e.g., those in SCOP fold a.118, that contain an up-and-down five-helix bundle, which is quite similar to the target structure except that the β-hairpin is replaced by another helix. This target was classified into the NF category because of the presence of the β-hairpin.

Since the hairpin (174–193) is only a small part of the whole domain, the models in which the hairpin is replaced by the helix of the template will still achieve a high GDT_TS score. We therefore visually inspected a large number of models, looking for those that had the hairpin. Most of the top scoring models have the helices modeled quite accurately, but have a helix at the place of the β-hairpin. Only five among the top 25 models had a nonhelix for residues (174–193). Models TS101_2 and TS450_2 are almost exactly the same (GDT scores 44.54 and 44.25) and have a hairpin for these residues. The hairpin is positioned well, but oriented about 10° off compared to that in the target structure. Model TS579_2u has a nonhelical structure for these residues also, but these form an impossible structure with many atomic clashes and chain overlaps. This is an unrefined model and it is interesting to note that the refined model TS579_2, the model with the best GDT score (50.0) for this target, has this part modeled as a nice helix. TS166_1 and TS021_2 also have a hairpin for the correct residues, but their orientations are off by approximately 20° and 30°, respectively.

T0272 [1WJ9 (1-85) and (90-211), FR/A, α + β]

The two domains of this target have a similar structure (Fig. 11). The core of each domain is a four-stranded β-sheet twisted around a main helix. The strands in the sheet are arranged in the order 4-1-3-2. The linker between β1 and β2 includes the main helix, which is α2 for the N-terminal domain (D1) and α1 in the C-terminal domain (D2). There are three helices in D1 and two in D2, only one of which, the main helix, is common. There are two missing segments in D2; one is between β1 and α1, which is helical in D1, and another between β2 and β3. The linker between β3 and β4 is irregular in D1 but helical (α2) in D2. The relation between the two domains in the whole structure is mainly translational, with a small of amount of rotation. The structure is related to d1lfwa2 and others in the SCOP Ferredoxin fold d.58.

Predictors could model D1 more accurately than D2. Both the average GDT_TS score (25.49 vs. 21.83) and the best GDT_TS score (58.53 vs. 34.59) are higher for D1 than for D2. Clearly this difference is not related to a difference in the topology, which is identical for the two domains. The difference could be related to the size difference since D1 with 85 residues is shorter than D2, which has 122 residues. One notes also that there are many missing residues in D2. Perhaps the missing parts have the sequences that make them structurally more flexible and difficult to model.

The three top-scoring models for D1 are all from Baker group (TS100_1, 2, and 3). They are all excellent (Fig. 11). The Cα RMSD between TS100_1 and D1 is only 3.7 Å for the entire domain of 85 residues. There is a large difference in quality between these models and the next best models, which is also reflected in the large difference in the

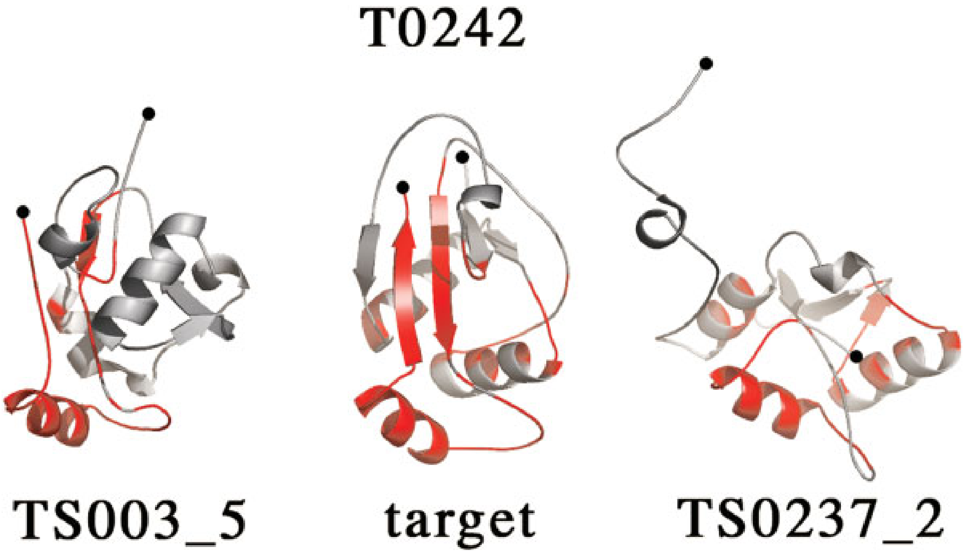

Fig. 10. Target T0242 aligned to models TS003_5 and TS237_2. The alignment, the secondary structure assignment, and the color scheme, are as described in Figure 2.

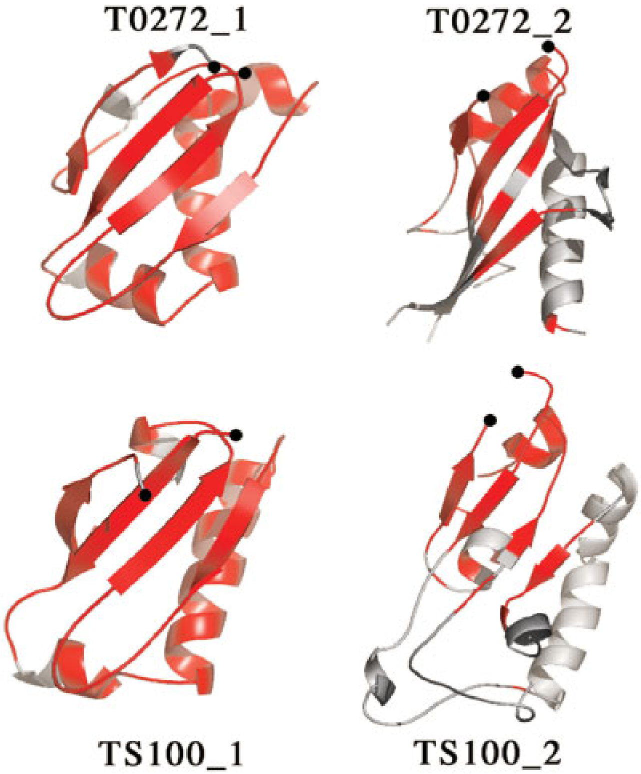

Fig. 11. Target T0272; domain 1 aligned to model TS100_1; domain 2 aligned to model TS100_2. The alignment, the secondary structure assignment, and the color scheme, are as described in Figure 2.

GDT_TS scores. Only two of the top-scoring 10 groups state the template used, which were not from the Ferredoxin fold.

The top-scoring two models for D2 are also from Baker group (TS100_2, and 1). The model for D2 in TS100_2 is poorer than that for D1 mainly because the model is inaccurate for the part between the C-terminal part of β1 through to the N-terminal part of β3. This region includes the two missing segments as well as the main helix and β2. The main helix and β2 are in the model but misplaced, presumably because the missing parts and some visible parts next to the missing parts are modeled incorrectly.

The model TS100_1 is similar to TS100_2 in that the same region between β1 and β3 is inaccurately modeled. In addition, the α2 between β3 and β4 is also misplaced in TS100_1 whereas it is quite accurately modeled in TS100_2. The GDT_TS score for this model is for an alignment that aligns the main helix accurately. This alignment aligns a few residues in α2 as well, but because the main helix is misplaced with respect to the sheet, this alignment does not match the β-sheet at all. The model can be realigned to match β1, β3, and β4 quite accurately. This alignment looks better because the β-sheet is aligned and the two helices are roughly in the correct region. However, LGA chose the α-helix-based alignment over this β-sheet-based alignment, presumably because the former produces more aligned residues than the latter. Similarly, the GDT_TS score for the fourth ranking TS100_5 is based mainly on the alignment of α1. This model is of comparable quality to TS100_1, except that β4 is not made. The third ranking model, TS579_2u, further demonstrates the "tyranny of α-helix." The GDT_TS score for this model is based entirely on α1 and the linker between α1 and β2. However, the model has a wrong domain boundary, lacks β-sheet in D2, and generally of rather poor quality.

T0273 [1WDJ (2–187), FR/A, α/β]

This structure is made of two subdomains (Fig. 12). D1 (2-41) is made of three antiparallel β-strands and a helix (α1). D2 (42–187), which forms the main bulk of the protein, has a three-layer structure. The central layer is a six-stranded β-sheet, on either side of which is a layer made of one helix and one β-hairpin loop. One of these

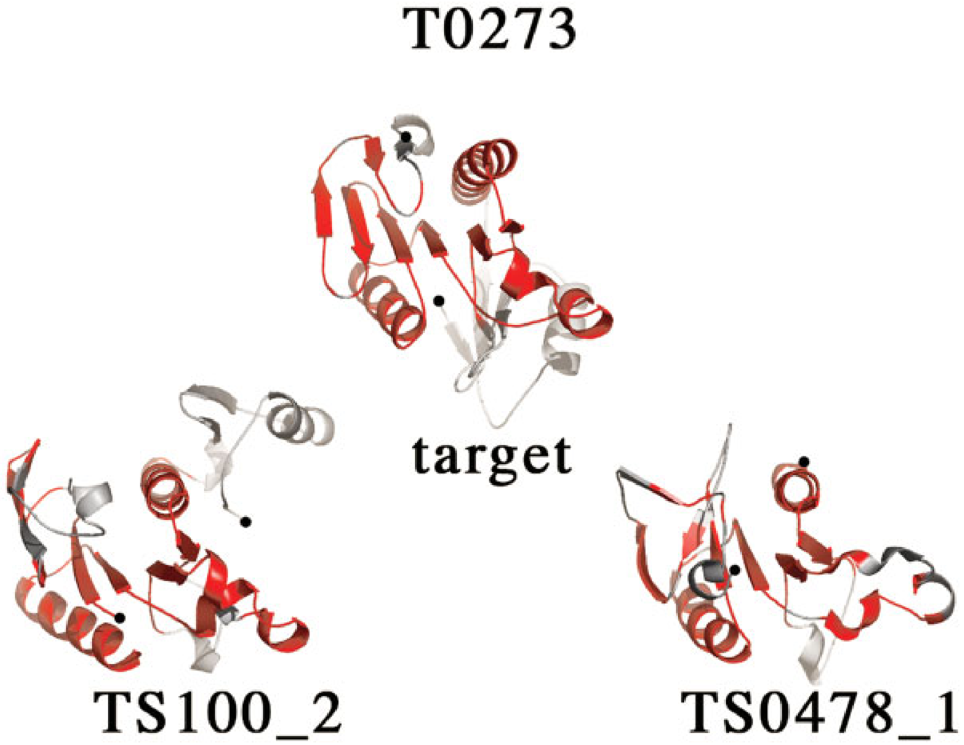

Fig. 12. Target T0273 aligned to models TS100_2 and TS478_1. The alignment, the secondary structure assignment, and the color scheme, are as described in Figure 2.

β-hairpins is between β4 and β7 and is oriented in such a fashion that it forms a continuous β-sheet with the three β-strands from D1. D2 clearly has the fold commonly found in restriction endonucleases, like d1fiua_ and d1gefa_ of the SCOP d.52.1 superfamily. The entire structure of D1 is contained in, for example, d1g51a2, which however is more than twice as large and contains other structures including a loop between the matching strands.

The top-scoring model is TS100_2 from the Baker group (Fig. 12). This is an excellent model with nearly perfect placement of all the major structural elements in both D1 and D2. The main defect, however, is the fact that D1 is placed on the opposite side of the central β-sheet and interacts with D2 through the α-helix (α2) rather than through the β-hairpin on the other side. The second- and third-best models are also from the Baker group (TS100_4 and TS100_1). These are of about the same quality as the top-scoring model and D1 is misplaced in a similar manner. The next set of eight high-scoring models is from just three groups, Boniaki-pred (TS478_1_2, Fig. 12, TS478_2_2), Kolinski&Bujnicki (TS021_4, TS021_3, TS021_1) and Ginalski (TS450_1, TS450_5, TS450_4). They all used d1gefa_ either solely or with other templates. These models are also good, but clearly have more defects than the Baker group's top-scoring models.

Summary of the Results of Visual Inspection

The best models, as judged by the visual inspection, are listed in Table I for each target. There was at least one model that had the correct fold for all but one (T0209_1) of the FR/A targets and for all three NFe targets. Two groups predicted the small key structural feature in T0248_2. The correct fold was not predicted for any of the six NFh targets.

Predictors were not particularly good at picking the best among their own predicted models: Among the 26 models listed in Table I, only eight are model one.

Table I shows that some group names occur more frequently than others. We counted the number of times a group's name occurs in the table for each category of targets. These counts are given in Table II (columns 3–5). Baker group was the most frequent producer of the best models; they submitted best models for 7 of the 17 targets. Five of these are for the FR/A targets and only two for the NF targets. The second-most frequent contributor of best models was Kolinski&Bujnicki group. Like Ginalski and Baker-Robetta groups, this group did best for the NFe targets. For the NFh targets, Baker and Jones-UCL groups were the most successful with two best models each, but five other groups produced best models for different targets in this category. Also, the ranking in this category is based on the quality of partial structures of models that do not have the overall fold of the target structure.

Ranking by Using the GDT_TS Scores

We ranked the performance of prediction groups objectively using GDT_TS scores and Z-scores for both model one and max model sets. There were four target sets: NFh, NFe, FR/A, and "all," which contains all the targets. Tables

TABLE II. Number of Times a Group Was Best or Among the Top Five in Different Target Categories

		Visually best[a]			Among top five[b]				Prediction counts[c]			
ID	Name	NFh	NFe	FR/A	NFh	NFe	FR/A	All	NFh	NFe	FR/A	All
100	Baker	1	0	5	4	2	4	4	6	3	8	17
021	Kolinski & Bujnicki	1	2	0	4	2	4	4	6	3	8	17
450	Ginalski	0	2	0	1	4	4	4	6	3	8	17
003	Jones-UCL	2	0	0	1	2	3	3	5	3	8	16
166	SAM-T04-Hand	—	—	—	2	4	0	2	6	3	8	17
604	Baker-Robetta_04	1	0	0	2	0	2	2	6	3	8	17
101	Baker-Robetta	0	2	0	2	2	0	1	6	3	8	17
160	Keasar	—	—	—	0	4	1	0	6	3	8	17
052	Rokky	—	—	—	2	0	0	0	6	3	8	17
060	Bilab	1	0	0	2	0	0	0	6	3	7	16
176	Skolnick-Zhang	—	—	—	0	0	2	0	6	3	8	17
501	Mcon	0	1	0	0	0	1	0	6	3	8	17
113	Pmodeller5	1	0	0	—	—	—	—	6	3	7	16
172	ProteinShop	1	0	0	—	—	—	—	2	3	6	11
035	GeneSilico	0	0	1	—	—	—	—	6	3	8	17
089	KIAS	0	0	1	—	—	—	—	6	3	8	17
157	3D-Jigsaw	0	0	1	—	—	—	—	6	3	8	17
185	Huber-Torda	0	0	1	—	—	—	—	4	3	8	15
530	B213-207	0	0	1	—	—	—	—	6	3	8	17

[a]Number of times a group submitted the best model by visual inspection for each target set.
[b]Number of times a group is in top five in Tables III–VI for each target set. The sum over all groups for each target set is sometimes >20 because of last place tie.
[c]Number of targets for which predictions were submitted by a group in each category.

TABLE III. Max Model Average GDT_TS Scores and Ranks for Each Target Set

Rank	NFh	NFe	FR/A	All
1	21.0 (021)[a]	54.5 (021)	39.9 (100)	34.4 (100)
2	21.0 (100)	48.8 (166)	30.9 (021)	31.6 (021)
3	20.7 (101)	48.6 (160)	29.9 (003)	29.4 (450)
4	20.3 (604)	48.5 (450)	29.3 (450)	28.5 (604)
5	19.9 (450)	48.1 (101)	29.2 (176)	28.3 (101)
6	19.7 (176)	46.1 (100)	28.2 (604)	28.2 (003)
7	19.6 (052)	45.7 (604)	27.5 (272)	28.1 (176)
8	19.5 (166)	45.0 (003)	27.1 (089)	27.5 (166)
9	19.4 (051)	42.0 (109)	27.1 (160)	27.1 (160)
10	19.3 (060)	41.8 (176)	27.1 (009)	26.6 (109)
11	19.3 (096)	41.2 (096)	26.8 (035)	26.2 (051)
12	19.1 (109)	39.8 (051)	26.6 (101)	26.0 (089)
13	18.9 (482)	39.5 (018)	26.6 (109)	25.5 (096)
14	18.8 (089)	39.0 (501)	26.1 (051)	25.0 (009)
15	18.2 (305)	38.9 (237)	25.8 (506)	25.0 (272)
16	18.0 (530)	38.6 (579)	25.7 (126)	24.7 (501)
17	17.9 (112)	38.0 (026)	25.6 (166)	24.7 (018)
18	17.9 (042)	37.4 (089)	25.4 (400)	24.7 (530)
19	17.8 (656)	36.7 (042)	25.4 (197)	24.4 (060)
20	17.8 (039)	36.4 (052)	25.4 (530)	24.3 (026)

[a]Numbers in parentheses are CASP group ID numbers.

TABLE IV. Max Model Average Z-Scores and Ranks for Each Target Set

Rank	NFh	NFe	FR/A	All
1	1.7 (021)[a]	2.6 (021)	2.8 (100)	2.2 (100)
2	1.5 (100)	1.9 (166)	1.6 (021)	1.8 (021)
3	1.5 (003)	1.9 (160)	1.3 (450)	1.4 (003)
4	1.5 (101)	1.8 (450)	1.3 (003)	1.3 (450)
5	1.3 (604)	1.8 (101)	1.1 (176)	1.3 (604)
6	1.2 (176)	1.6 (100)	1.1 (604)	1.2 (101)
7	1.1 (052)	1.5 (003)	1.0 (089)	1.2 (176)
8	1.1 (450)	1.5 (604)	1.0 (035)	1.0 (166)
9	1.0 (060)	1.2 (109)	0.9 (009)	0.9 (051)
10	1.0 (166)	1.2 (176)	0.9 (051)	0.9 (089)
11	0.9 (051)	1.0 (096)	0.9 (101)	0.8 (109)
12	0.9 (096)	1.0 (579)	0.9 (272)	0.8 (060)
13	0.9 (482)	1.0 (051)	0.8 (060)	0.8 (160)
14	0.9 (109)	0.9 (501)	0.8 (160)	0.8 (096)
15	0.8 (478)	0.9 (018)	0.8 (381)	0.8 (035)
16	0.8 (089)	0.8 (237)	0.8 (094)	0.7 (501)
17	0.7 (113)	0.8 (094)	0.7 (166)	0.7 (052)
18	0.7 (112)	0.7 (530)	0.7 (109)	0.7 (009)
19	0.7 (501)	0.7 (026)	0.7 (400)	0.6 (272)
20	0.6 (035)	0.6 (033)	0.7 (185)	0.6 (018)

[a]Numbers in parentheses are CASP group ID numbers.

III and IV give the average GDT_TS and Z-scores, respectively, for the top 20 groups in each target set. These are for the max model set. Similar tables for the model one set are given in the Tables V and VI.

Average GDT_TS scores are noticeably lower for the model-one set than for the max model set indicating that predictors did not always choose the best model as their model one. This was also evident from visual inspection of individual top-scoring models (see above). SAM-T04-hand group (166) was the best among top scoring groups in selecting the best model one, as indicated by the fact that their ranking substantially increased when model one set is used.

The largest average Z-score for the NFh targets is only 1.7 for the best model set (Table IV). The corresponding values for the NFe and FR/A sets are 2.6 and 2.8. This

TABLE V. Model One Average GDT_TS Scores and Ranks for Each Target Set

Rank	NFh	NFe	FR/A	All
1	19.0 (166)[a]	47.3 (166)	36.8 (100)	31.8 (100)
2	18.3 (100)	45.6 (100)	27.7 (021)	26.4 (021)
3	18.2 (021)	44.1 (160)	27.6 (450)	26.2 (450)
4	18.2 (052)	44.1 (003)	25.4 (003)	26.1 (166)
5	18.1 (060)	42.1 (450)	25.3 (160)	24.9 (003)
6	17.6 (039)	41.7 (604)	25.0 (109)	24.8 (160)
7	17.5 (176)	39.4 (051)	24.9 (501)	24.7 (501)
8	17.4 (400)	39.1 (021)	24.7 (506)	24.6 (604)
9	17.3 (348)	39.0 (101)	24.6 (176)	24.3 (051)
10	17.3 (051)	39.0 (501)	24.5 (604)	24.3 (101)
11	17.3 (501)	37.0 (176)	24.1 (454)	24.3 (176)
12	17.2 (113)	36.7 (089)	24.1 (101)	23.8 (109)
13	17.2 (305)	36.4 (018)	23.9 (051)	23.3 (089)
14	17.1 (101)	35.9 (109)	23.6 (272)	22.5 (506)
15	17.1 (478)	35.6 (579)	23.5 (166)	22.5 (454)
16	16.5 (114)	34.9 (344)	23.5 (089)	22.5 (018)
17	16.5 (042)	34.8 (052)	23.4 (035)	21.7 (400)
18	16.5 (450)	34.8 (042)	23.3 (504)	21.7 (344)
19	16.4 (089)	33.6 (506)	22.9 (007)	21.6 (052)
20	16.1 (604)	33.0 (007)	22.8 (011)	21.5 (035)

[a]Numbers in parentheses are CASP group ID numbers.

TABLE VI. Model One Average Z-Scores and Ranks for Each Target Set

Rank	NF	NFe	FR/A	All
1	1.4 (166)[a]	2.3 (166)	2.8 (100)	2.1 (100)
2	1.3 (021)	2.0 (100)	1.5 (021)	1.4 (021)
3	1.3 (052)	1.9 (003)	1.4 (450)	1.2 (166)
4	1.2 (100)	1.8 (160)	1.0 (501)	1.2 (450)
5	1.1 (060)	1.5 (450)	1.0 (604)	1.1 (003)
6	1.1 (478)	1.5 (604)	0.9 (160)	1.0 (501)
7	1.0 (501)	1.3 (021)	0.9 (101)	1.0 (101)
8	1.0 (113)	1.3 (051)	0.9 (109)	0.9 (176)
9	1.0 (039)	1.1 (101)	0.9 (003)	0.9 (604)
10	1.0 (101)	1.1 (501)	0.9 (051)	0.9 (051)
11	1.0 (176)	1.0 (579)	0.9 (176)	0.8 (160)
12	0.9 (003)	1.0 (176)	0.8 (506)	0.8 (109)
13	0.9 (348)	1.0 (094)	0.8 (454)	0.7 (060)
14	0.8 (051)	1.0 (018)	0.8 (089)	0.7 (089)
15	0.8 (305)	0.9 (089)	0.8 (166)	0.7 (113)
16	0.8 (289)	0.9 (109)	0.8 (113)	0.6 (052)
17	0.8 (400)	0.9 (352)	0.7 (272)	0.6 (018)
18	0.7 (064)	0.9 (573)	0.7 (060)	0.6 (454)
19	0.7 (450)	0.8 (406)	0.7 (035)	0.6 (042)
20	0.7 (019)	0.8 (207)	0.7 (011)	0.6 (348)

[a]Numbers in parentheses are CASP group ID numbers.

means that the difference between the best and the average models are not very large for the NFh targets. The closeness of GDT_TS values for the NFh models make the ranking among these models unstable, particularly when the score drops even a little below the very best. For these difficult targets, even a large difference in ranking probably means little since most models are poor.

Tables III–VI show that some groups populate the top ranges of all categories more frequently than others. To summarize the performance of each group, we counted the number of times a group was within the top five in the four tables for each target set. This count is given in the last four columns of Table II. (The rankings reported here are somewhat different from those presented at the December CASP6 meeting for three reasons: we excluded FR/A targets with average GDT_TS score < 25 and grouped the included targets differently in order to more cleanly separate the NFh and FR/A targets; we dropped the average rank category as it gives too much penalty for nonsubmissions; and we used a newer raw dataset, which does not include a number of student submissions that were included in the old dataset.)

The two groups most often in the top five, Baker (100) and Kolinski&Bujnicki (021) groups, have identical top-five membership profiles. Table III shows that Baker group did better for the FR/A targets but more poorly for the NFe targets. For the NFh targets, they were even (GDT_TS score) or worse (Z-score) than the Kolinski&Bujnicki group. The NFh average is better for the latter group mainly because this group produced an exceptionally good model for T0216_1 compared to all other models (Fig. 5). For other NFh targets, scores are nearly the same or clearly better for the Baker group (data not shown).

The NFh average GDT_TS score is low for Jones-UCL group because of one nonsubmission; they did not submit a prediction for T0238. If T0238 is excluded, this group has the best averages for the NFh targets (data not shown).

Comparison with the Visual Inspection Results

A number of groups produced a best model by visual inspection for only one target but are not listed in any top five GDT_TS score counts (Table II). Presumably these groups are not consistent in their predictions so that their average score for a given set of targets is not among the top five. Four groups did well on average for a given target set and are listed in the top five counts but did not produce a best model for any category. All five groups who produced more than one visually best model are ranked in the top five counts.

DISCUSSION

We had nine NF targets this year compared to five in CASP5. Many are hypothetical proteins of unknown function from the structural genomics project.[14] For example, T0216 is a large two-domain protein of 426 residues (Fig. 4) in which both domains have folds that have not been seen before and for which the function is unknown. T0241 is also a two-domain protein of new fold (Fig. 4) from the structural genomics project. It is a little smaller (237 residues) but still large for an NF target. As new-fold structure prediction becomes difficult when the structure is large, the prediction community had a particularly strong challenge this year.

In the December 2004 CASP6 meeting, we presented the evaluation results for all targets in both NF and FR/A categories. We include for this report only the eight most difficult FR/A targets (those for which the average GDT_TS score was at or below 25) in addition to the nine NF targets. We also grouped these targets into three nonoverlapping groups to more cleanly separate the NF easy (NFe) and difficult (NFh) targets from the FR/A targets.

The evaluation was done in two complementary ways. Best models for each target were selected from visual inspection of models that had high GDT_TS scores. For numerical ranking we used the GDT_TS scores averaged over different target sets.

GDT_TS Scores and Quality of Models

We found, by visual inspection of many models, that GDT_TS scores usually provide a reliable measure of model quality. However, as already described by the CASP5 NF target evaluators,[4] there were exceptional cases for which GDT_TS scores presented a problem. Aloy et al. pointed out three types of problems: (1) When the target structure contains a prominent helix, models with a long helix get high GDT_TS scores regardless of the rest of the structure; (2) for α/β structures, the GDT_TS score is rather insensitive to small errors in the position (order) of β-strands in the sheet; and (3) when the model contains segments that overlap, the GDT_TS score can be elevated because some of the 4 and 8A component of the score derives from the overlapping segment. We have also seen all of these cases as described below.

The cases wherein a prominent helix in the target dictated the GDT_TS score occur for three targets. For targets T0216_2 and T0238, this is the reason why the models judged to be the best are not those with the highest GDT_TS score (Table I). However, these are difficult targets, for which even the best models have only a part of the structure correctly modeled. Also, the visually picked best models and highest scoring models differ by less than 1 GDT_TS unit. For the FR/A target T0272_2, the helix effect did elevate an incorrect model to the position of rank 3. However, this did not affect the overall ranking since the group that submitted this model (579) was not among the top 20 groups in terms of the average score for the FR/A targets (Table III).

The errors of β-strand position in high scoring models for the α/β structures have been noted in the Results section for targets T0201 and T0216_1. However, these are still the best models. We have not encountered an example wherein a better model was ranked lower because of the insensitivity of GDT_TS score with respect to the strand positioning.

There were some high-scoring models that had serious Cα overlaps because some of the loops were misplaced. But we felt that most that we have seen were errors that can be easily fixed by a rather simple refinement of the structure that preserves overall topology of the model. Unlike CM models, for example, wherein such errors should probably be treated seriously, we felt that these errors were relatively minor in view of the generally low overall quality of the models. Bad models might have benefited a little by such an error, but it was not clear that the GDT_TS of the good models would in fact be lowered had the model been refined. We included them without change in calculating the GDT_TS averages.

We also spotted what must be a rare example wherein a poor model had a high score mainly because it had three spatially separated helices in approximately correct relative positions. (See the Results section for target T0242 and Fig. 10.)

In summary, we believe that the averages given in Tables III–VI, which are based on the GDT_TS scores, do provide a fairly true reflection of the average quality of the models each prediction group generated. It is obviously desirable to have a scoring scheme that does not have these defects, but we suspect that other scoring systems will have their own idiosyncrasies and that it is not an easy task to find a perfect scoring system.

Nonsubmissions

When comparing the performance of predicting groups, the method for handling nonsubmissions is an issue. Different groups may choose not to submit predictions for certain targets for different reasons. For instance, group A may decide not to submit a model for difficult targets while group B may decide not to compete for easy targets for which templates are available.

One way of handling nonsubmissions is to ignore them and use the score averaged over the submitted predictions only. This can result in group A of above example scoring better than group B even when B is better than A for common targets. Another method is to compare groups using only the common targets, i.e., those targets for which all the groups being compared made predictions (head-to-head comparison). However, this can result in a large number of combinations of groups, since different groups may submit for different sets of targets. Also, the results of such head-to-head comparisons are generally non-transitive, i.e., if A is better than B in one head-to-head comparison and B is better than C in another, A is not necessarily better than C in the head-to-head comparison of A and C.

Instead, we chose to treat nonsubmissions as predictions with an arbitrarily assigned score of zero for both the GDT_TS score and the Z-score. Since it does not seem fair that a predicted model should score lower than a nonsubmission, we also set all negative Z-scores to zero. Setting all negative Z-scores to zero has the effect of ignoring score differences among below-average models. This is not necessarily an undesirable effect since numerical score differences often mean little when the predicted models are very different from the target structure. It should also be noted that the impact of a nonsubmission is greater for the GDT_TS score than for the Z-score, because GDT_TS scores are larger in absolute value than Z-scores so that a score of zero is relatively smaller when compared to the GDT_TS scores than to the Z-scores.

Other Ranking Issues

For the quantitative ranking, we used both the GDT_TS and Z-scores averaged over each target set. As described above, nonsubmissions and negative Z-scores were both set to zero before taking the averages. The two averages are complementary to each other. Generally, GDT_TS scores are preferable since they provide an absolute measure of model quality. The Z-score measures the quality of a model relative to other models and, in principle, even a poor model can have a high Z-score as long as other models are even poorer. Fortunately, this does not happen in practice. For difficult targets for which even the best model is poor, the spread between the best and the average GDT_TS scores is small (see below) and the Z-score of the best model is small. (See the Z-scores for the NFh targets in Table IV.) Also, nonsubmissions have a much greater effect on the GDT_TS average than on the Z-score average, as pointed out above and as we saw in the Results section in the case of the Jones-UCL group.

The top five membership count is a simple device for combining the four different averages given in Tables III–VI. It is blind to small rank differences within top five. It is also an elitist measure that ignores less than outstanding performers.

One additional caveat may be mentioned: The rankings presented here are based on a relatively small sample of only 17 targets. Although some groups are clearly better than others for the current set of targets, the small sample size implies that the conclusion may not be statistically significant and that they might not do well for the next batch of new targets. Only when the group performs consistently well with many different target sets, can we be sure, in the statistical sense, that the method they employ is indeed better.

Quality of the Predictions

As is evident from Tables I–III, the Baker group was clearly the most successful for the FR/A targets. For five targets they were best by a large margin; the difference between the best and second-best ranged from 5–14 in the GDT_TS scale. The average Z-score for the Baker group was about 80% better than that of the second best (Tables III and VI). For the α-helical bundle T0198, the β-sandwich T0212, and the α/β ferredoxin-like T0272_1, their best models are nearly flawless.

For the NFh targets, the Baker group was not as successful. Table II shows that seven different groups produced one or more best models in this category. However, even the best models have only parts of the structure correctly modeled. In these instances it is difficult to judge one partial structure as being better than another partial structure. Generally, the quality of models in this category is rather disappointing, except perhaps for T0216_1.

All three NFe targets have substructures that match a substructure in other proteins and a large number of groups produced good to excellent models for all three. Target T0248_2 is a special case; the structure is like an up-and-down five-helix bundle, for which many known structures exist, except that one of the helices is replaced by a β-hairpin. The question was whether this small but conspicuous feature of the structure could be predicted. As expected, most predictions had a helix at the place of the β-hairpin. However, a model from the Baker-Robetta group and a very similar one from the Ginalski group did have the β-hairpin, although not positioned quite correctly.

Blurring of NF and FR Techniques

Many predicting groups now use both de novo and homology modeling/fold recognition techniques to predict structures in all categories. CASP experiments themselves, which tend to reward all-target predictors, may have encouraged this trend. In any case, it makes sense for a laboratory to have both techniques available. A homology modeling/fold recognition task requires a de novo technique in order to model the loops and other unaligned parts. Even when one is primarily interested in a de novo technique, one has to first recognize which targets/domains are good NF candidate and which not. The best way to find out is to try homology modeling/fold recognition first. All four top groups listed in Table II as well as the two servers from Baker's laboratory use both methods.[15–19] For example, the Baker-Robetta_04 group (604) used homology modeling to produce some models for a given target and a de novo protocol for others for the same target. For the NFh targets T0241_1 and T0238, a de novo built model was their best for one and a homology built model was their best for the other (private communication). Since the introduction of the fragment assembly technique by Jones[20] and Baker's group[21] in 1997, the technique has become quite popular; most successful prediction groups now use some form of fragment assembly as a part of their de novo structure prediction protocol. It was evident at CASP5[4,22] and now also in CASP6 that the fragment assembly technique is highly successful, and effectively competes with the FR technique, at least for some of the more difficult FR/A targets. For the assessors, this means that it is impossible to determine the type of technique used from the character of the target alone.

As far as we are aware, most groups use these two techniques independently or sequentially, with little interaction between them. Meta-servers like Mcon (501) provide an extreme example wherein the two steps are completely separated but used sequentially. Mcon[13] takes server models, some of which are produced using a de novo procedure by another server, and identifies the best among them (fold recognition). However, the Kolinski&Bujnicki group used a new hybrid technique in which homology models from Frankenstein's Monster[23] and de novo models from CAFASP servers[24] were used to obtain spatial restraints that guided the next de novo folding step.[16]They produced an excellent model for the NF target T0201 and the best model for the NF target T0216_1. Since these are targets with no good template, and since their models are substantially better than the next best model, these are presumably produced mainly by the de novo part of their technique. On the other hand, they also submitted one of the four best and identical models for T0209_2, presumably by recognizing a server model (from Baker-Robetta, 101) as being the correct model (fold recognition).

The Baker group added a new protocol: for some models they obtained β-strand pairing information from models from a meta-server (bioinfo.pl) and then used this information as constraints in their de novo ROSETTA fragment assembly algorithm.[15] They produced excellent models for the two FR/A targets, T0212 and T0273 using this protocol (personal communication), which were much better than other models produced using a template in both cases. While this protocol does not use spatial constraints from the models, the β-strand pairings should be powerful constraints for these β-sandwich (T0212) and three-layer β-sheet (T0273) structures.

Although these two techniques differ greatly in detail, they are also similar in that both use a de novo folding method that is guided (constrained/restrained) by information obtained from other models, which were themselves built using a fold recognition/homology modeling procedure (or another de novo procedure or a previous round of the same procedure). Such a technique may normally be considered a de novo method except that, when there is a sufficiently good model in the pool of models that provide guidance, a good system may simply produce the constraining model or one very similar to it and behave much like a fold recognition system.

Why Are Some Groups More Successful Than Others?

The four top scoring models for T209_2 were identical. Presumably three of these are copies of the model originally made by the Baker-Robetta server. The copiers deserve credit for recognizing this as a good model, since they did not adopt the much poorer model for the first domain of the same protein from the same server. Another example occurs for T0248_2. The model by Ginalski group is nearly identical to that from Baker-Robetta server.

Obviously, there is the problem of assigning credit for the original prediction. It is also difficult to call the technique NF even though the target is. Other than these bookkeeping issues, however, it is obviously better to recognize a good model from a server than to produce a wrong model. The advent of the hybrid techniques means that many models may start from a copy of a server model and may undergo a substantial modification before becoming the final model. The examples cited above are noticeable because they are exact or nearly exact copies of other models. It is possible that there are other near copies that we have not noticed.

We detect that there is a more subtle form of copying happening among the best techniques. We note from Table I and Figure 1 that some of the FR/A targets are more "difficult", i.e., have lower average GDT_TS score, than some NFh targets, yet the best predictors were more successful with the former than the latter. Indeed, the spread between the best and the average scores is larger for the FR/A targets than for the NFh targets. Since closely related structures exist for the FR/A targets, this suggests that the most successful predictors could exploit the information from similar structures more skillfully than others. When closely similar structures do not exist, as for the NFh targets, this skill is less useful and the spread decreases.

The disappointing results for NFh targets suggest that the prediction community as a whole has learned to copy well but has not really learned how proteins fold. This is not necessarily alarming. As structural genomics projects progress, more new folds will be determined by experimental techniques and less new folds left undiscovered. From a practical point of view, learning to copy well will be increasingly more important than being able to predict a new fold.

Recommendations for the Future

The following two areas need to be considered with some urgency for future CASP experiment. One has to do with assessment categories and the other with the model quality estimate.

The assessment in the NF category is to assess the prediction techniques that do not use a template. However, as described above, it is now difficult or impossible to tell, from the target alone or even when the identity of the predicting group is known, if a particular prediction was made using or not using a template. As a consequence, we had a great difficulty in deciding which targets to include for evaluation in the NF category. One way to solve this problem is to simply draw an arbitrary line, as we have done, on the basis of the average GDT_TS score. Another possibility would be to ask the predictors to specify, for each model they submit, the prediction category in which they wish their models to be evaluated.

There is a large difference in quality between the best NFh models and the best FR/A models and between the best and average models for all targets. We believe that it is nearly irresponsible for the prediction community to present these different quality models to a nonpredicting scientific colleague without an assessment of the quality of the models. In order to encourage predictors to assign confidence measures to their predictions, we suggest that the predictors be asked to submit their predicted GDT_TS score (or other common quality measure). The predicted score can be evaluated along with the model quality or even as a separate prediction category.

ACKNOWLEDGMENT

We thank the CASP organizers for making this experience possible. It has been an honor and pleasure to serve as one of the assessing groups. We particularly thank Andriy Kryshtafovych, Anna Tramontano, and John Moult, without whom this work may never have been finished. We thank the other assessors, Roland Dunbrack, Michael Tress, and Alfonso Valencia, for the numerous discussions and suggestions throughout the assessment process. We thank the predictors for their countless hours of work in this CASP experiment. This research was supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

REFERENCES

1. http://predictioncenter.org/casp6/.
2. Tress M, Tai C-H, Wang G, Ezkurdia I, López G, Valencia A, Lee B-K, Dunbrack RL Jr. Domain definition and target classification for CASP6. Proteins 2005;Suppl 7:8–18.

3. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
4. Aloy P, Stark A, Hadley C, Russell R. Predictions without templates: new folds, secondary structure, and contacts in CASP5. Proteins 2003;Suppl 6:436–456.
5. DeLano WL. The PyMOL molecular graphics system. http://www.pymol.org.
6. Chandonia JM, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. ASTRAL compendium enhancements. Nucleic Acids Res 2002;30:260–263.
7. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL Compendium in 2004. Nucleic Acids Res 2004;32(Database issue):D189–192.
8. Jung J, Lee B. Protein structure alignment using environmental profiles. Protein Eng 2000;13:535–543.
9. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
10. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res 1998;26:320–322.
11. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2002;30:276–280.
12. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32(Database issue):D226–229.
13. Azaria Y. Selecting models with a meta-MQAP. CASP6 Meeting Abstracts 2004:97; http://predictioncenter.org/casp6/abstracts.
14. Zhang C, Kim SH. Overview of structural genomics: from structure to function. Curr Opin Chem Biol 2003;7:28–32.
15. Bradley P, Cheng G, Chivian D, Kim D, Malmstrom L, Meiler J, Misura K, Qian J, Schonbrun J, Zanghellini A, et al. Novel approaches to protein structure prediction at CASP6. CASP6 Meeting Abstracts 2004:9–11. http://predictioncenter.org/casp6/abstracts.
16. Koliński A, Bujnicki JM. Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. CASP6 Meeting Abstracts 2004:88–89. http://predictioncenter.org/casp6/abstracts.
17. Ginalski K. Modeling of CASP6 target proteins with 3D-Jury, Meta-BASIC and ROSETTA. CASP6 Meeting Abstracts 2004:64. http://predictioncenter.org/casp6/abstracts.
18. Sadowski MI, Watson JD, Sodhi JS, Ward JJ, Jones DT. FRAGFOLD3, THREADER3 and DISOPRED2: improved methods for prediction of protein folds, disorder and function. CASP6 Meeting Abstracts 2004:81–82. http://predictioncenter.org/casp6/abstracts.
19. Chivian D, Kim DE, Malmstrom L, Schonbrun J, Rohl CA, Baker D. The Robetta and Robetta_04 protocols. CASP6 Meeting Abstracts 2004:11–12. http://predictioncenter.org/casp6/abstracts.
20. Jones DT. Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. Proteins 1997;Suppl 1:185–191.
21. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
22. Kinch LN, Wrabl JO, Krishna SS, Majumdar I, Sadreyev RI, Qi Y, Pei J, Cheng H, Grishin NV. CASP5 assessment of fold recognition target predictions. Proteins 2003;Suppl 6:395–409.
23. Kosinski J, Cymerman IA, Feder M, Kurowski MA, Sasin JM, Bujnicki JM. A "FRankenstein's monster" approach to comparative modeling: merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation. Proteins 2003;Suppl 6:369–379.
24. Fischer D, Rychlewski L, Dunbrack Jr. RL, Ortiz AR, Elofsson A. CAFASP3: the third critical assessment of fully automated structure prediction methods. Proteins 2003;Suppl 6:503–516.