PROTEINS: Structure, Function, and Genetics 53:395–409 (2003)

CASP5 Assessment of Fold Recognition Target Predictions

Lisa N. Kinch,[2]* James O. Wrabl,[2] S. Sri Krishna,[1] Indraneel Majumdar,[1] Ruslan I. Sadreyev,[2] Yuan Qi,[1] Jimin Pei,[1] Hua Cheng,[1] and Nick V. Grishin[1,2]

[1]*Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas*

[2] *Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas*

ABSTRACT **We present an overview of the fifth round of Critical Assessment of Protein Structure Prediction (CASP5) fold recognition category. Prediction models were evaluated by using six different structural measures and four different alignment measures, and these scores were compared to those assigned manually over a diverse subset of target domains. Scores were combined to compare overall performance of participating groups and to estimate rank significance. The methods used by a few groups outperformed all other methods in terms of the evaluated criteria and could be considered state-of-the-art in structure prediction. We discuss a few examples of difficult fold recognition targets to highlight the progress of ab initio-type methods on difficult structure analogs and the difficulties of predicting multidomain targets and selecting prediction models. We also compared the results of manual groups to those of automatic servers evaluated in parallel by CAFASP, showing that the top performing automated server structure predictions approached those of the best manual predictors. Proteins 2003;53:395–409.** © 2003 Wiley-Liss, Inc.

Key words: protein fold prediction; structure comparison; alignment quality; threading; domain structure; CASP5

INTRODUCTION

We present a detailed report of the CASP5 fold recognition assessment, which included evaluating prediction models, quantifying these evaluations in a meaningful way, and using these measurements to produce a ranking of groups that was subject to various tests of significance. Automation played a crucial role in completion of the analysis. Our report outlines the resulting evaluation procedure, the logic behind its development, and the results of its application to CASP5 fold recognition target predictions. On the basis of these results, we highlight the pitfalls and progress of both the top performing prediction groups and the fold prediction community as a whole.

In addition to defining CASP5 target domain classifications (see Target Classification article in this issue[1]), the assessment of the CASP5 fold recognition category encompassed evaluations of both the structural quality of model predictions with respect to defined experimental targets (see Table I) and the alignment quality derived from such predictions. To help with this task, the Livermore Prediction Center provided automated evaluations of submitted model predictions using methods developed throughout the decade-long course of CASP.[2–6] Such measures proved to be an essential component of the assessment, considering the growing number of experimental targets (46 fold recognition domains) and the increasing number of groups represented in CASP5 (149 manual groups and 59 automatic servers predicted fold recognition domains). For the fold recognition category alone, 20220 predictions had to be evaluated, making manual judgment of all predictions impossible within the required timescale of the assessment.

Given the drawbacks of relying on a single measure to estimate the quality of all fold recognition predictions (e.g., see Fig. 3), we decided to incorporate scores from recognized structural comparison methods including Dali,[7,8] CE,[9] and a relatively new structural comparison method, Mammoth.[10] In addition, we incorporated scores from a contact distance method we developed for the purpose of CASP evaluation (see Methods and Results). We considered scores from these four methods, along with the two Livermore automated evaluation measures (GDT_TS and SOV_O scores[3–6]) to reflect the overall structural quality of the predictions.

In previous CASP assessments of the fold recognition category, alignment quality had been an important component of evaluation.[11–14] To expand on this design, we chose to evaluate the quality of prediction alignments using the output from structural superpositions of DALI, CE, and Mammoth, in addition to the output from a sequence-independent structure alignment (LGA[5]) provided by the group at Livermore. The quality of alignments produced by these four structural superposition methods was scored independently on the basis of the fraction of correctly aligned residues. Thus, our overall evaluation included scores generated by six different structural measures and four different alignment measures for every fold recognition target prediction.

To compare the overall prediction quality of different groups, scores generated with all measures and for all fold recognition domains had to be combined to produce a single value reflective of group performance, despite the fact that groups predicted varying numbers of targets.

*Correspondence to: Lisa N. Kinch, Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390-9050. E-mail: lkinch@chop.swmed.edu.

Received 28 February 2003; Accepted 17 June 2003

TABLE I. Fold Recognition Domains

Target[a]	Domain borders[b]	Class[c]
T0130	6-105	CM/FR(H)
T0132	B6-B152	CM/FR(H)
T0133	A16-A308	CM/FR(H)
T0134_1	A878-A1006	FR(H)
T0134_2	A1007-A1112	FR(H)
T0135	C3-C108	FR(A)
T0136_1	E4-E259	CM/FR(H)
T0136_2	E260-E523	CM/FR(H)
T0138	A1-A135	FR(H)
T0141	A1-A187	CM/FR(H)
T0146_1	A1-A24, A114-A196	FR(A)/NF
T0146_2	A25-A113	FR(A)/NF
T0146_3	A244-A299	FR(A)/NF
T0147	A2-A245	FR(A)
T0148_1	A2-A9, A101-A163	FR(A)
T0148_2	A10-A100	FR(A)
T0149_1	A2-A202	CM/FR(H)
T0152	G12-G209	CM/FR(H)
T0156	A2-A157	FR(H)
T0157	A2-A138	FR(H)
T0159_1	X1-X91, X234-X309	CM/FR(H)
T0159_2	X92-X233	CM/FR(H)
T0162_1	A7-A62	FR(A)
T0162_2	A63-A113	FR(A)
T0165	A1-A318	CM/FR(H)
T0168_1	A1-A68, A210-A327	CM/FR(H)
T0168_2	A69-A209	CM/FR(H)
T0169	A1-A156	CM/FR(H)
T0170	1-69	FR(A)/NF
T0172_1	A2-A115, A217-A294	CM/FR(H)
T0172_2	A116-A216	FR(A)/NF
T0173	A3-A299	FR(A)/NF
T0174_1	B8-B28, B199-B374	FR(H)
T0174_2	B39-B198	FR(H)
T0184_2	B166-B236	CM/FR(H)
T0185_1	A1-A101	CM/FR(H)
T0185_3	A299-A446	CM/FR(H)
T0186_2	A45-A256, A293-A330	CM/FR(H)
T0186_3	A257-A292	FR(A)/NF
T0187_1	A4-A22, A250-A417	FR(A)/NF
T0187_2	A23-A249	FR(A)
T0189	A1-A319	CM/FR(H)
T0191_1	A1-A104, A248-A282	FR(A)
T0192	A2-A153, B154-B171	CM/FR(H)
T0193_1	A1-A78	FR(H)
T0195	A1-A299	CM/FR(H)

[a]CASP5 target codes. Domain numbers are appended to the codes for multidomain proteins.
[b]Domain boundaries according to Experimental model residue numbers.
[c]Target class. CM: Comparative Modeling; FR: Fold Recognition; (H): Homologous; (A): Analogous; NF: New Fold.

Given that each measure provided different scores and that each target varied in difficulty, any procedure used to combine scores had to include rescaling. To address this problem, we chose to assign scores (z scores) based on a comparison with the average prediction scores for individual target domains. Once scaled according to each target, group scores were combined by using various strategies and averaged over all measures to produce values reflective of the overall performance of each group.

This general procedure allowed us to compare and rank each group participating in CASP5 fold recognition prediction and to assign statistical significance to the results. Although group rankings were based entirely on these automatic scores, we believed that including some aspect of manual judgment of predictions would add value to the assessment. In addition to manually inspecting a number of models for each target domain, we chose a subset of 10-fold recognition domains that ranged in difficulty to score manually (scoring all prediction models with correct overall fold; see Methods and Results) and compared these results to our automatic scoring method. To gain insight into which groups performed better on different types of targets, we evaluated subsets of fold recognition homologues and analogs separately. To assess which groups performed better at different tasks, we evaluated structural quality measures and alignment quality measures separately. These various subsets of scores helped us to more closely evaluate the most successful approaches to fold recognition.

METHODS AND RESULTS

Fold Recognition Target Domains

Traditionally, the fold recognition category has included domains that fell in between the more clearly defined comparative modeling targets (displayed sequence similarity to known folds) and the new fold targets (displayed no structural similarity to known folds), with some overlap. Because of the increasing power of sequence similarity detection methods, the overlap between comparative modeling and fold recognition domains in CASP5 has now become extensive and difficult to define. Previous CASP domain classifications have been rather subjective and limited to available sequence (defined by PSI-BLAST[15,16] in CASP4) and structural (defined by ProSup[17] in CASP4) information for similar folds.[11,13] We chose to expand on these traditional domain class boundary definitions by combining a broader measure of sequence similarity with an objective estimation of target difficulty based on the quality of submitted predictions.

Target sequence similarity to the closest known fold can be established by a variety of methods. As measured by PSI-BLAST procedures, similarity estimates depend on the program version, sequence length, database size, and input cutoffs.[15,16] A measure that does not depend on the structure of sequence space or on arbitrary cutoff values would provide a more reliable estimate of similarity. Starting with a structure-based alignment to the closest PDB template, we estimated sequence similarity with the following score (S_{seq}):

$$S_{\mathrm{seq}} = \frac{S_{12} - S(\mathrm{random})}{S(\mathrm{max}) - S(\mathrm{random})} = \frac{S_{12} - \sum_{i,j}^{20} f_i^{(1)} f_j^{(2)} s_{ij}}{\frac{S_{11} + S_{22}}{2} - \sum_{i,j}^{20} f_i^{(1)} f_j^{(2)} s_{ij}}$$

where S_{12} represents the score between the aligned target (1) and template (2) sequence measured by the BLOSUM62 similarity matrix ignoring gap regions, S_{11} and

S_{22} represent the same scores between the target (1) sequence and itself and the template (2) sequence and itself ignoring gap regions, s_{ij} represents the same score between amino acids i and j, and $f_i^{(1)}$ and $f_j^{(2)}$ represent the frequency of amino acid i in the target sequence (1) and the frequency of amino acid j in the template sequence (2), respectively. This score provided an estimate of sequence similarity that did not depend on the structure of sequence space. Furthermore, this score correlated with sequence identity between all CASP5 domain target/template pairs ($r = 0.974$) but was more sensitive to target/template pair differences at low-sequence identity.

Although similarity measures indicate the general predictability of targets, the performance of the predictor community as a whole provides the most direct estimate of target difficulty. As expected, the target similarities estimated by S_{seq} scores generally correlated with target difficulties estimated by average GDT_TS scores [diagonal line, Fig. 1(A)]. A model-based clustering[18] of this data suggested the existence of four different groups and reflected the discrete nature of protein space. We chose to use this clustering to define the boundary between the fold recognition assessment [red targets, some blue targets, Fig. 1(A)] and the comparative modeling assessment [purple and green targets, Fig. 1(A)]. The relatively low average GDT_TS scores and similarity scores of the red and blue clusters suggested differences in overall fold that warranted evaluation using fold recognition assessment criteria. Because most of the domain sequences from the red cluster [Fig. 1(A)] found known folds with sequence-based methods, this boundary definition resulted in an extensive overlap between the two assessment categories [21 CM/FR(H) domains].

Unfortunately, these clusters became less meaningful for defining the boundary between new fold and fold recognition targets. The structure templates required to estimate similarity (S_{seq} score) were somewhat arbitrary for new folds and difficult fold recognition domains. Thus, to establish this boundary, we used traditional structural criteria (see classification article for discussion[1]). The cluster reflecting the most difficult domains (blue, Fig. 1) contained all but two (T0170 T0172_2) of the resulting new folds (NF and FR(A)/NF domains). Table I summarizes all domains considered in the CASP 5 fold recognition category and defines the respective domain borders according to residue numbers of experimental models.

Domains included in the fold recognition category could be split into homologues (30 domains) and analogs (16 domains) of known folds. We briefly summarize the classification illustrated in Table I: comparative modeling/fold recognition domains that detected known folds using sequence-based methods were considered to be homologues [CM/FR(H)]. This overlapping category included 22 domains, which represented almost half of the total number of identified fold recognition targets (46 domains). Fold recognition domains that did not detect known folds with sequence-based methods but possess substantial structural similarities to known folds were also considered to be homologues [FR(H)]. For the eight domains of this class, various justifications allowed an inference of evolutionary relatedness (see target classification article[1]). Fold recognition analog domains [FR(A)] contained general structural similarity to known folds, although evolutionary relatedness could not be established. Finally, new fold/fold recognition (NF/FR) domains contained very distant structural similarities to known folds. For these eight domains, topologies generally resembled those of known folds, but the overall size or packing arrangement of secondary structural element differed significantly.

Numerical Evaluation of Predictions

To evaluate the overall quality of fold recognition predictions for a given target, each prediction must be compared with the experimental model by considering both structural similarities and alignment quality. Ideally, an overall score could be assigned to each prediction based on these criteria. Unfortunately, no single standard measure exists to generate such scores, despite an expanding and diverse number of available protein structure comparison methods (for a review, see Ref. 19). To aid in our assessment, the group at Livermore provided various evaluation scores (GDT_TS, SOV_O, and LGA_Q, e.g.) including three structural superpositions (2 sequence dependent and 1 sequence independent) for each prediction ranked by target in order of descending GDT_TS scores.[20] Given the possible limitations of using a single measure to estimate the overall quality of structural predictions and the diverse range of fold recognition targets, we decided to apply a combination of available structural comparison measures to our assessment of fold recognition domains. Just as different predictions captured various details of fold recognition targets, different structural comparison methods potentially captured various aspects of such predictions and provided a certain robustness to method-specific errors.

For our evaluation, we chose to use the global distance test total score (GDT_TS) and the segment overlap measure observed score (SOV_O), which have been developed over the time course of CASP evaluation.[2–4,6,21] The GDT_TS score represents the average of the percentage of residues that can be superimposed within a given distance over four optimal sequence-dependent superpositions (1, 2, 4, and 8 Å).[2,4,6] This GDT analysis has been generally accepted by the prediction community and has provided the basis for previous CASP assessments. Although the GDT_TS score reflects the overall tertiary structural quality of a model prediction, the SOV_O score provides an assessment of prediction quality that depends on a segment-based evaluation of secondary structure.[3,21] These two scores provided alternative automated evaluations of CASP predictions that addressed both global (GDT_TS) and local (SOV_O) aspects of model quality.

To further increase the range of model evaluation, we chose to include scores from three additional structural evaluation methods developed in other laboratories whose programs were available for local use (Dali,[7,8] http://www.ebi.ac.uk/dali/; CE,[9] http://cl.sdsc.edu/ce.html; and Mammoth,[10] http://icb.mssm.edu/services/mammoth/mam-

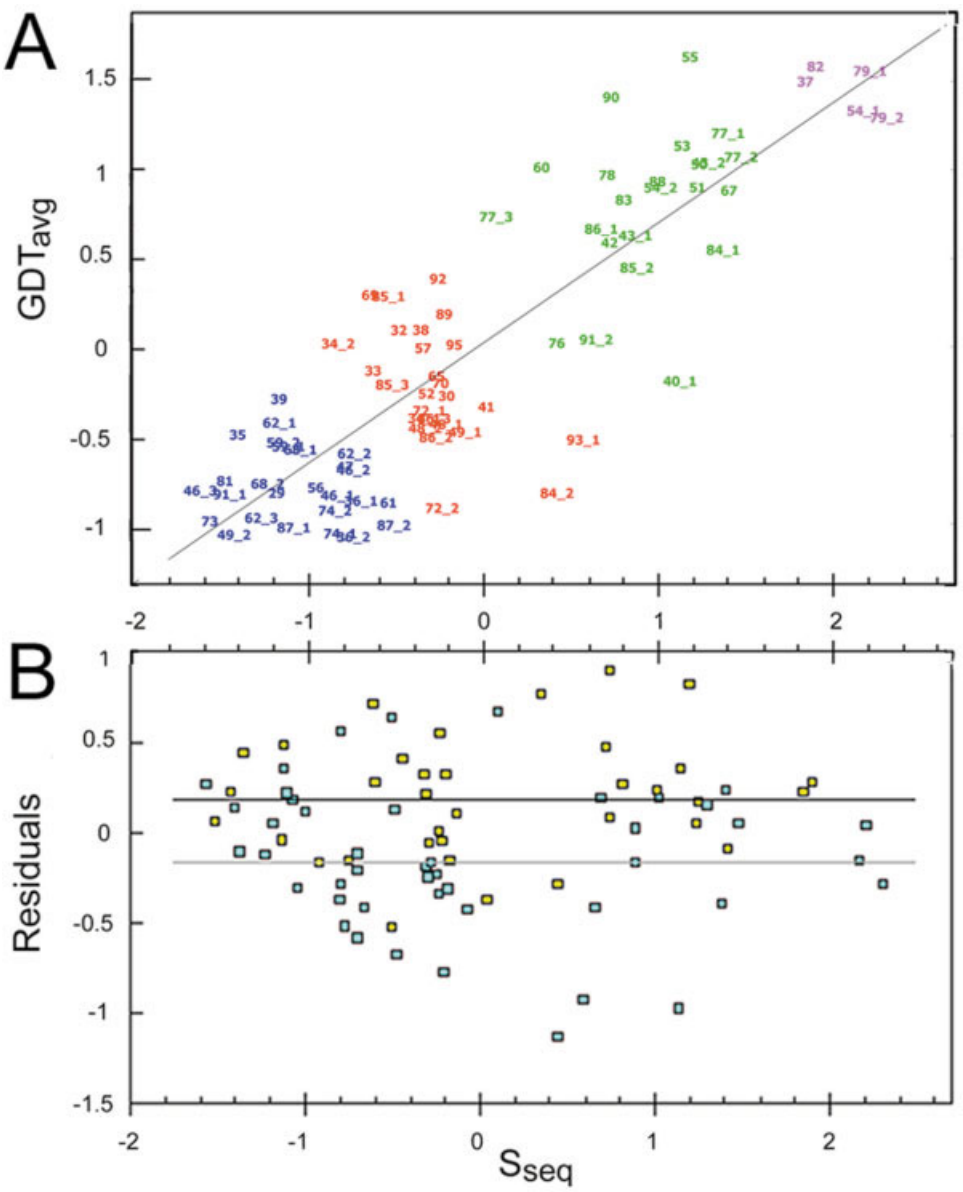

Fig. 1. Target difficulty and domain classification. **A:** Average GDT_TS scores for all target domains rescaled according to *z* score (GDT_TS_{avg}) are plotted against BLOSUM (S_{seq}) similarity scores generated from a structure-based sequence alignment of the target with the closest template and rescaled by *z* score. The linear regression fit (diagonal gray line) represents predicted average GDT_TS scores for difficulty levels represented by S_{seq}. Each data point is labeled according to domain and colored according to cluster analysis.[18] For clarity, we omitted the "T01" from the start of all target identification numbers so that T0129 became 29, for example. **B:** The residuals of average GDT_TS scores with respect to predicted average GDT_TS scores from data shown in (A) are plotted against the same similarity score described in (A). Data from single-domain targets are colored yellow, and data from multiple domains are colored light blue. A black line and a gray line indicate the averages of all single-domain and multiple-domain residuals, respectively.

moth). Of these structural alignment programs, both Dali and CE compare intramolecular Cα geometries (distances and angles, respectively) of the target structure with those of the model structure. Although the two programs use different procedures to generate optimal structural alignments, each defines the quality of the results in terms of a *z* score. Results of the two methods tend to diverge significantly for dissimilar structures (SCOP superfamily/family pairs under about 10% sequence identity)[22] and may thus detect different aspects of structural predictions for the various fold recognition targets with marginal predictions.

The third structural evaluation method (Mammoth) was developed to evaluate structural similarities at the fold level, which seems ideal for scoring domains that fall within the fold recognition category. Mammoth differs from the previous two methods by producing structural alignments that do not depend on contact maps but instead depend on unit-vector root-mean-square dis-

tances. The program also associates a statistical significance value to the similarity score.[10] Such a method allows for possible registration shifts that occur frequently in structural analogs and concentrates on structural similarities at a more general level than does Dali or CE.

As an additional structural evaluation method, we generated a sequence-dependent score that relied on similarities between intramolecular Cα contact distances of the target and model structures. The contact distance score was calculated with the following equation:

$$S_{\text{contact}} = \left[\sum_{i=1}^{N} \sum_{j=i+1}^{N} e^{[-(dij - di'j')^2/\sigma]} \right]$$

where *dij* represented the distance between the Cα atoms of residues *i* and *j* in the target structure, *di'j'* represented corresponding residues *i'* and *j'* in the model structure, *N* represented the total number of amino acid residues in the target structure, and σ represented an adjustable parameter that was taken to be the value (0.5 Å^2) at which the correlation between GDT_TS and contact score was maximized over all models for all targets. LiveBench uses a similar measure to continuously evaluate structure prediction servers.[23]

To evaluate the alignment quality of fold recognition predictions, we used sequence alignments produced by the sequence-independent structural superposition methods [Dali, CE, Mammoth, and the local-global alignment (LGA) provided by Livermore]. Alignment quality (Q) scores for each of the different superpositions was defined as the fraction of correctly aligned residues. Raw scores were generated by adding one point for each correctly aligned residue from the structural superposition and dividing the sum by the length of the target sequence. No fractional points are awarded for shifts. These scores were rather stringent and could be viewed as structure-independent measures of alignment quality.

Combining Scores

With a goal of comparing the overall prediction quality of different groups, scores generated by using the six structure and four alignment measures needed to be combined to produce a single value. The fact that each target exhibited a different level of difficulty complicated this task, making combined scores meaningful only after proper scaling across targets had been applied. Over the course of CASP, different assessors have approached this scaling problem in different ways. For example, in the CASP4 fold recognition assessment, scores were normalized by assigning an overall numerical value (1–4) to each prediction based on manually assigned thresholds that were different for each target.[11] We decided to apply a less subjective methodology that rescaled prediction scores according to the average prediction quality of that target (using *z* scores). For each participating group, these *z* scores could be either summed or averaged over all targets (or a subset of targets) and then averaged over all measures (or a subset of measures).

As a first approach to combine target scores, we chose to apply the following equation (*z* score summation) in which the overall score for a group *j* is defined as:

$$score\ j(\text{sum}) = \sum_{i=1}^{N} \left[\frac{score\ i - \langle score\ i \rangle_{\text{predictions}}}{\sigma_{i,\text{predictions}}} \right]$$

where *N* represents the number of predicted FR domains, $\langle score\ i \rangle_{\text{predictions}}$ represents the average score of all groups for a given target *i* with a given measure, and $\sigma_{i,\text{predictions}}$ represents the standard deviation. This summation produced a single group score for each given measure. By summing the scores of each target domain, groups who predicted more targets had an opportunity to achieve a higher overall score than those that predicted fewer targets (although penalties for worse than average predictions counted as negative *z* scores). Thus, as a second approach to combining group scores, we decided to average *z* scores over all predicted targets.

$$score\ j(\text{mean}) = \sum_{i=1}^{N} \left[\frac{score\ i - \langle score\ i \rangle_{\text{predictions},2\sigma}}{N \cdot \sigma_{i,\text{predictions},2\sigma}} \right]$$

For this method, we recalculated the mean ($\langle score\ i \rangle_{\text{predictions},2\sigma}$) and sigma ($\sigma_{i,\text{predictions},2\sigma}$) disregarding scores below $-2*\sigma$ of the entire sample, and we divided by the number of predicted domains (N) to achieve an average. In the process of averaging, each negative *z* score was replaced with zero to eliminate the penalty for worse than average predictions. This method of combining scores was based on the same principles outlined in current and previous assessments of comparative modeling targets.[24,25] After combining the target scores, each group was assigned a single summation score [score j(sum)] and a single average score [score j(mean)] for each measure (6 structural measures and 4 alignment measures). Again, we rescaled both the summation scores and the average scores for each measure to obtain *z* scores.

Once rescaled, *z* scores for each individual measure could either be compared with the *z* scores produced by the other measures directly (Fig. 2) or could be combined by simple averaging to generate overall scores (Tables II–IV scores). Linear correlation coefficients (Pearson *r* values) from pairwise plots of *z* scores for individual measures (first models) are reported in Figure 2. The *z* scores of each individual measure compared to those averaged over all measures suggested that as a single measure, the GDT_TS score ($r = 0.973$) most closely represented that of all combined measures, with Mammoth following closely behind ($r = 0.964$).

Although correlations between all measures suggested that the GDT_TS score was the single most representative measure for assessing fold recognition targets, the GDT_TS score alone fell short in evaluations of difficult targets. For one such target [T0147, Fig. 3(A)], GDT_TS scores placed at rank 4 a structural prediction [Fig. 3(B)] that appeared fragmented, included some overlapping coordinates, and contained irregular helices and strands. Alternatively,

	GDT	Contact	DALI	CE	MM	SOV	Q_{lga}	Q_{dali}	Q_{mm}	Q_{ce}	Avg.
GDT	1	0.931	0.838	0.834	0.937	0.86	0.929	0.916	0.874	0.862	0.973
Contact		1	0.699	0.752	0.824	0.876	0.792	0.784	0.855	0.823	0.903
DALI			1	0.909	0.898	0.752	0.861	0.891	0.613	0.775	0.892
CE				1	0.887	0.815	0.822	0.855	0.69	0.882	0.914
MM					1	0.777	0.943	0.942	0.824	0.867	0.964
SOV						1	0.717	0.755	0.767	0.796	0.879
Q_{lga}							1	0.97	0.836	0.865	0.946
Q_{dali}								1	0.827	0.881	0.955
Q_{mm}									1	0.84	0.88
Q_{ce}										1	0.93
Avg.											1

Fig. 2. Measure correlations. *z* scores using first models of all fold recognition targets for each measure were plotted against those of every other measure and the average scores over all measures (Avg). The Pearson *r* value for each comparison is reported for pairwise measure comparisons. Individual measures are abbreviated as GDT_TS for GDT_TS structural measure, Contact for contact distance measure, DALI for Dali structural measure, CE for CE structural measure, MM for Mammoth structural measure, SOV_O for segment overlap measure, Q_{lga} for LGA superposition alignment score, Q_{dali} for Dali alignment, Q_{ce} for CE alignment, and Q_{mm} for mammoth alignment.

TABLE II. Top 20 Predictors Ranked by Combined Scores, All Domains

Rank first (sum)[a]	Rank best (sum)[a]	Rank first (mean)[b,c]	Rank best (mean)[b]	Group[c,d]	Predictor[c,d,e]	Predictions scored	*z* score first (sum)[a,d]	Bootstrap first (sum)[a,d]	*z* score best (sum)[a,d]	Bootstrap best (sum)[a,d]
1	1	2	1	**2**	**Baker**	46	**3.21**	**0.83**	**3.58**	**0.98**
2	3	3	6	**453**	**Ginalski**	46	**2.76**	**0.70**	1.87	0.26
3	8	6	13	**6**	**Rychlewski**	46	**2.37**	**0.62**	1.60	0.10
4	6	10	12	29	Robetta (S)	46	2.02	0.33	1.68	0.12
5	13	7	14	517	Bujnicki	43	1.96	0.29	1.33	0.04
6	2	12	3	**10**	**Skolnick**	46	1.88	0.17	**2.69**	**0.97**
7	25	14	32	96	Bates	46	1.66	0.12	0.93	0.07
8	18	17	23	427	Fischer	46	1.64	0.18	1.18	0.07
9	11	4	5	20	Bujnicki	33	1.61	0.13	1.38	0.17
10	23	13	21	110	Honig	36	1.50	0.09	1.05	0.09
11	16	15	16	28	Shi	39	1.49	0.05	1.20	0.05
12	5	22	9	12	Xu	46	1.46	0.08	1.79	0.23
13	14	19	22	450	Labesse	42	1.43	0.05	1.25	0.07
14	9	27	19	373	Brooks	46	1.35	0.07	1.48	0.14
15	31	18	28	153	Takeda-Shitaka	39	1.34	0.10	0.83	0.04
16	17	31	29	112	Friesner	46	1.24	0.05	1.19	0.07
16	45	32	72	67	Jones	46	1.24	0.07	0.54	0.05
18	20	29	31	1	Karplus	46	1.21	0.07	1.14	0.09
18	7	23	10	40	Pmodel (S)	46	1.21	0.06	1.65	0.14
20	3	21	8	45	Pmodel3 (S)	46	1.15	0.04	1.87	0.38

[a]Sum refers to combining target scores by summation as described in the text.
[b]Mean refers to combining target scores by averaging *z* scores (-2σ, no negative scores) as described in the text.
[c]Group 425 (Venclovas) and group 448 (Murzin) do not appear in the top 20 using the summation method (predicted 7 domains and 22 domains, respectively). Venclovas ranked 1 and Murzin ranked 5 using averaging method of combining scores.
[d]Groups/scores in bold display significant (>95% confidence) difference from most of the remaining top 20 groups using Student's *t*-tests.
[e]Automatic servers are indicated with (S) following the predictor name.

GDT_TS scores ranked at 27 a structural prediction [Fig. 3(C)] with a general fold topology identical to the target. Manual inspection of these two structures disagreed with these ranks. For this target, our combined score strategy caused the fragmented prediction to drop to rank 70 (due to unfavorable structural measure scores), whereas the topologically correct prediction moved to a higher rank (10). Although such extreme cases of GDT_TS shortcomings occur infrequently, this example highlights the danger of relying on a simple ranking scheme for evaluation. Using the input from more than one evaluation method provides one way to increase the significance of scores and thus the meaning of ranks, because the shortcomings of individual methods are essentially averaged over the inputs of other methods.

Our evaluation process assigned a single overall score to each prediction group based on the input of 10 different automated scores. How did this process compare to a manual evaluation? The large number of fold recognition predictions to be evaluated (20220 models) prevented a complete comparison of manual scores with those generated automatically. Therefore, we chose a subset of 10 fold recognition domains (T0130, T0132, T0133, T0135, T0138, T0141, T0147, T0156, T0157, and T0173) for manual evaluation to address this question. The chosen domains represented a broad range of difficulty levels (average

TABLE III. Top 20 Predictors Ranked by Combined Scores, Homologues, and Analogs

Homologues						Analogs					
Rank first (sum)[a]	Rank best (sum)[a]	Group[b]	Predictor[b,c]	Predictions scored	*z* score first (sum)[a,b]	Rank first (sum)[a]	Rank best (sum)[a]	Group[b]	Predictor[b,c]	Predictions scored	*z* score first (sum)[a,b]
1	3	453	**Ginalski**	30	**2.56**	1	1	**2**	**Baker**	16	**5.09**
2	5	6	**Rychlewski**	30	**2.27**	2	6	**349**	**Shortle**	12	**2.73**
3	8	517	**Bujnicki**	30	**2.00**	3	4	**1**	**Karplus**	16	**2.23**
4	1	2	**Baker**	30	**1.95**	4	21	453	Ginalski	16	2.14
5	12	96	**Bates**	30	**1.88**	5	3	29	Robetta (S)	16	2.08
6	10	20	Bujnicki	27	1.88	6	31	6	Rychlewski	16	1.69
7	9	427	Fischer	30	1.74	7	2	10	Skolnick	16	1.65
8	16	29	Robetta (S)	30	1.65	8	7	68	Jones	8	1.47
9	2	10	Skolnick	30	1.63	9	17	51	Samudrala	16	1.26
10	18	110	Honig	27	1.50	10	51	67	Jones	16	1.17
11	15	28	Shi	30	1.49	10	14	112	Friesner	16	1.17
12	6	12	Xu	30	1.47	12	59	132	I-sites/Bystroff (S)	16	1.09
13	22	153	Takeda-Shitaka	28	1.46	13	43	517	Bujnicki	13	1.03
14	13	450	Labesse	28	1.40	14	15	373	Brooks	16	0.94
15	7	40	Pmodel (S)	30	1.33	15	34	28	Shi	9	0.91
16	27	368	Saldanha	28	1.30	16	45	110	Honig	9	0.89
17	34	265	Sasson-Iris	30	1.27	17	29	450	Labesse	14	0.88
18	11	373	Brooks	30	1.26	18	64	423	Taylor	16	0.79
19	28	84	Elofsson	30	1.22	19	8	16	Levitt	14	0.78
20	26	448	Murzin	17	1.21	19	9	12	Xu	16	0.78

[a]Sum refers to combining target scores by summation as described in the text.
[b]Groups/scores in bold display significant (>95% confidence) difference from most of the remaining top twenty groups using Student's *t*-tests.
[c]Automatic servers are indicated with (S) following the predictor name.

TABLE IV. Top 20 Predictors Ranked by Structure Measures and Alignment Measures

Structure						Alignment					
Rank first (sum)[a]	Rank best (sum)[a]	Group[b]	Predictor[b,c]	Predictions scored	*z* score first (sum)[a,b]	Rank first (sum)[a]	Rank best (sum)[a]	Group[b]	Predictor[b,c]	Predictions scored	*z* score first (sum)[a,b]
1	3	**453**	**Ginalski**	25	**3.11**	1	1	**2**	**Baker**	25	**3.54**
2	1	**2**	**Baker**	25	**2.74**	2	5	**453**	**Ginalski**	25	**3.00**
3	5	**6**	**Rychlewski**	25	**2.57**	3	3	**29**	**Robetta (S)**	25	**2.84**
4	2	10	Skolnick	25	1.97	4	7	**6**	**Rychlewski**	25	**2.76**
5	11	**110**	**Honig**	20	**1.87**	5	2	10	Skolnick	25	2.69
6	9	29	Robetta (S)	25	1.78	6	14	427	Fischer	25	1.90
7	22	517	Bujnicki	25	1.70	7	15	28	Shi	23	1.88
8	32	96	Bates	25	1.54	8	35	96	Bates	25	1.67
9	21	427	Fischer	25	1.51	9	24	517	Bujnicki	25	1.64
10	15	28	Shi	23	1.45	10	9	1	Karplus	25	1.63
11	7	450	Labesse	22	1.44	11	27	110	Honig	20	1.61
12	19	20	Bujnicki	17	1.32	12	8	373	Brooks	25	1.60
13	26	448	Murzin	15	1.30	13	32	448	Murzin	15	1.38
14	47	67	Jones	25	1.27	14	26	20	Bujnicki	17	1.32
15	9	41	Akiyama	25	1.25	15	6	12	Xu	25	1.30
16	37	153	Takeda-Shitaka	20	1.22	16	55	67	Jones	25	1.20
17	39	435	Fujita	21	1.20	17	53	265	Sasson-Iris	25	1.15
18	57	105	Sternberg	25	1.15	17	17	450	Labesse	22	1.15
19	49	368	Saldanha	23	1.14	19	47	242	Bujnicki	22	1.05
20	13	373	Brooks	25	1.12	20	21	68	Jones	10	1.03

[a]Sum refers to combining target scores by summation as described in the text.
[b]Groups/scores in bold display significant (>95% confidence) difference from most of the remaining top 20 groups using Student's *t*-tests.
[c]Automatic servers are indicated with (S) following the predictor name.

GDT_TS scores from 14.0 to 49.4) and included both homologues and analogs of known folds.

To construct manual scores, one point was assigned to each correctly predicted secondary structural element, and one point was added for every secondary structural element that produced correctly aligned residues. Half points were subtracted for erroneous distances, gaps, or non-protein-like architectures. Based on the individual tar-

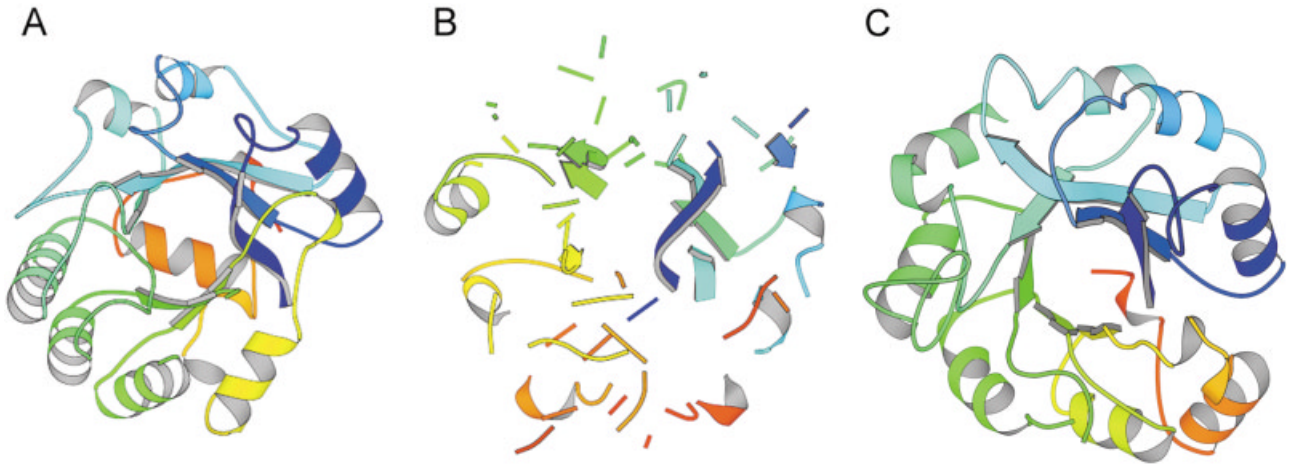

Fig. 3. Example of GDT_TS measure shortcoming. **A:** Target 147 *E. coli* YcdX experimental model. **B:** Example of GDT_TS score 32.8 (rank 4) prediction. **C:** Example of GDT_TS score 27.57 (rank 27) prediction. Objects are colored by using a rainbow ramp of secondary structure progression from N-terminal (blue) to C-terminal (red). Figures are prepared with MolScript.[34]

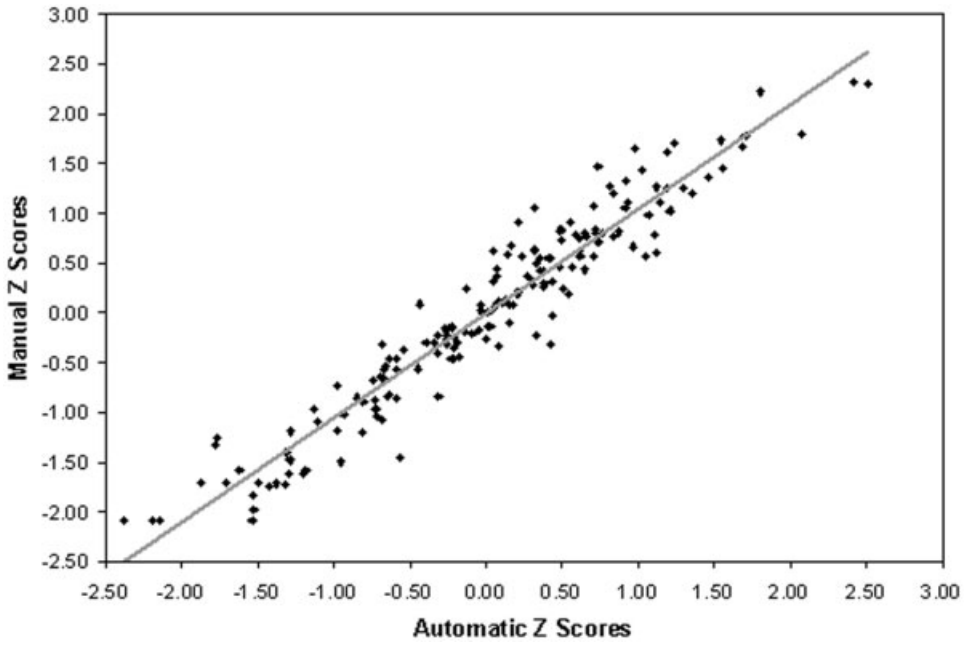

Fig. 4. Comparison of manual evaluation scores to automated evaluation scores. Manual evaluation scores were generated as described in the text. These manual scores (*y* axis) were summed over all evaluated targets (10 domains) and rescaled according to *z* score. The automatic scores (*x* axis) were generated by averaging *z* scores generated by the six structural and four alignment measures over the same targets (10 domains). The data points represent scores for each first model prediction for all evaluated targets, and the gray trend line fit by linear regression shows the correlation between scores generated manually and automatically ($r = 0.96$).

gets, various bonus points were also possible for functional or conserved residue placement or for unusual structural feature predictions. All models for each target were evaluated in order of decreasing GDT_TS score, and evaluations stopped when predictions no longer contained a majority of the target fold. Thus, the number of evaluations ranged from 74 different models for T0173 (difficult target) to 200 different models for T0141 (easy target). When the unevaluated models were assigned a score of zero, the resulting manually assigned scores correlated with the averaged scores generated for the same subset of target domains using our automated procedure ($r = 0.960$, first models, Fig. 4).

Predictors were allowed to submit up to five models for each target. Although this evaluation focused on models labeled "1" (first models), the organizers of CASP did not outline specific criteria for addressing the additional models. In the evaluation of the fold recognition category for

	2	453	6	29	517	10	96	427	20	110	28	12	450	373	153	112	67	1	40	45
2	0	0.9	1.71	3.05	2.39	2.77	3.02	3.34	1.32	2.91	2.78	3.86	3.39	3.77	3.02	4.48	4.31	5.29	4.14	4.18
453	46	0	2.86	2.03	3.16	3.17	3.94	3.62	2.41	4.32	3.98	4.41	4.61	4.72	3.87	5.07	5.34	5.06	5.03	4.13
6	46	46	0	0.99	1.49	1.88	2.81	2.54	1.11	2.56	2.93	2.89	2.99	3.62	2.52	3.63	4.42	3.58	3.88	3.29
29	46	46	46	0	-0.1	0.43	1	1.33	-1.3	0.34	0.76	1.73	1.29	1.86	1.54	2.26	2.61	2.59	2.38	2.55
517	43	43	43	43	0	0.39	0.89	1.36	1.09	1.74	1.55	2.15	1.69	1.64	2.53	2.63	2.34	2.71	2.48	2.28
10	46	46	46	46	43	0	0.6	0.72	-0.7	-0.2	0.69	1.15	0.79	1.73	0.8	1.74	2.04	2	1.79	1.73
96	46	46	46	46	43	46	0	0.12	-0.8	-0.2	0.8	0.73	0.53	1.06	0.14	1.27	1.52	1.12	1.61	1.66
427	46	46	46	46	43	46	46	0	-1.6	0.05	0.32	0.65	0.27	0.75	1.1	1.5	1.46	1.22	1.51	1.23
20	33	33	33	33	33	33	33	33	0	1.44	2.89	2.7	1.8	1.22	2.54	3	2.12	3.06	2.18	1.62
110	36	36	36	36	36	36	36	36	30	0	-0	0.52	0.36	2.84	0.9	1.58	1.63	2.49	1.15	2.47
28	39	39	39	39	39	39	39	39	33	36	0	0.44	-0.3	0.67	0.53	1.47	1.15	1.65	0.9	1.39
12	46	46	46	46	43	46	46	46	33	36	39	0	-0.4	0.31	-0.4	0.81	0.79	0.69	0.97	0.87
450	42	42	42	42	39	42	42	42	33	34	37	42	0	0.49	0.32	1.33	0.88	0.95	1.28	0.72
373	46	46	46	46	43	46	46	46	33	36	39	46	42	0	-1.9	0.28	0.39	0.38	0.43	0.75
153	39	39	39	39	36	39	39	39	31	35	36	39	37	39	0	0.86	0.75	0.98	0.9	2.09
112	46	46	46	46	43	46	46	46	33	36	39	46	42	46	39	0	0.05	0.09	0.14	0.25
67	46	46	46	46	43	46	46	46	33	36	39	46	42	46	39	46	0	0.03	0.09	0.24
1	46	46	46	46	43	46	46	46	33	36	39	46	42	46	39	46	46	0	0.05	0.19
40	46	46	46	46	43	46	46	46	33	36	39	46	42	46	39	46	46	46	0	0.2
45	46	46	46	46	43	46	46	46	33	36	39	46	42	46	39	46	46	46	46	0

Fig. 5. Statistical significance by Student's *t*-test. The number of domains common to both groups are listed to the left of the diagonal, whereas the values of *t* are listed to the right of the diagonal. Statistically significant *t* values that correspond to a probability of deviation >95% are highlighted in gray.

CASP4, only first models were considered. For our assessment, we chose to evaluate both "first models" and "best models" to ascertain the best-quality models produced by given methods. To give the predictors every opportunity to obtain a higher overall score, we defined the best model as the model with the highest score for a given measure. This best score may not necessarily represent the same model number for all of the measures (on average, ~7.5 best models were identical out of 10 measures). Overall rankings appeared to be quite sensitive to whether first or best models were used as input. For example, the automatic server Pmodel3 ranked 3 when using best models of all fold recognition domains but dropped to rank 20 when considering first models (see Table II).

Evaluation of Significance

Table II summarizes the top 20 prediction groups ranked by "first" model combined measure scores for all defined fold recognition domains (46 domains). By using the summation method, the highest scores achieved by group 2 (Baker, 3.21), group 453 (Ginalski, 2.72), and group 6 (Rychlewski, 2.37) appeared to stand apart from the remaining scores (<2.02). To test if the prediction quality of these groups could be reliably distinguished from those of the other groups, we sought to evaluate the statistical significance of the results using various measures. Previously, the statistical significance between paired samples of CASP4 comparative model domains were estimated by using the parametric Student's *t*-test and the nonparametric Wilcoxon signed rank test.[26] Figure 5 illustrates the comparison of the top 20 CASP5 fold recognition groups using the paired Student's *t*-test on combined measure scores for all fold recognition targets (first models). Indeed, these three groups tended to perform significantly differently from the remaining groups (gray highlights > 95% confidence of difference). Below these top groups, the statistical significance of the differences in ranking was marginal. Results were similar for evaluations using the Wilcoxon signed rank test (data not shown).

Successful application of Student's *t*-test to target predictions assumes a normal distribution of model quality differences (although this is not an assumption of the Wilcoxon test) and a large number of common models.[11,27] Although most of the top scoring groups predicted most or all of the fold recognition targets, these assumptions did not necessarily hold for all groups participating in CASP. Therefore, we applied an additional evaluation of rank significance to the fold recognition results through a bootstrap selection of *z* scores.[27] In the bootstrap procedure, the score for each group over all targets was calculated from a random selection of N z scores, where N equals the number of targets predicted by the group and the selection set was composed of the actual first or best model *z* scores. The selection was performed N times with returns. Thus, some *z* scores may have been selected more than once, whereas others may not have been represented at all. The scores obtained from this random selection process were used to rerank the groups. This procedure was repeated 200 times, and the number of times each group obtained the same bootstrap rank as their actual calculated rank was tabulated and reported as a fraction of the total number of repeats (Table II). By using this procedure, a larger bootstrap value at a particular rank indicated a higher significance of that rank. The results of this bootstrap procedure also suggested that the ranks of the top three groups were significantly better than those of the other groups.

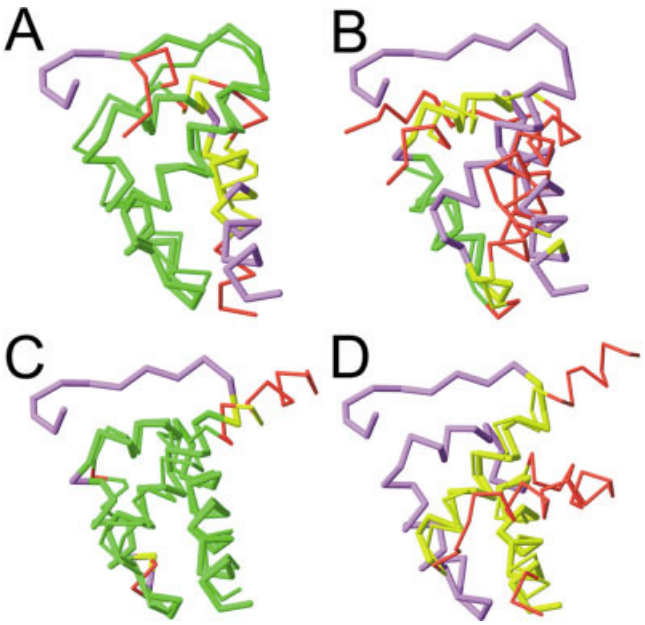

Fig. 6. First models and best models. Target T0170 sequence-independent structural superpositions (Livermore) with (**A**) group 2 Baker prediction best model labeled "4," (**B**) group 2 Baker prediction first model, (**C**) group 10 Skolnick prediction best model labeled "3," and (**D**) group 10 Skolnick prediction first model. In all superpositions, the experimental model structure is colored in thick purple, and the prediction is colored in red. If the distance between aligned residues is below 4 Å, then the residue backbone traces are colored yellow (not aligned correctly) or green (aligned correctly).

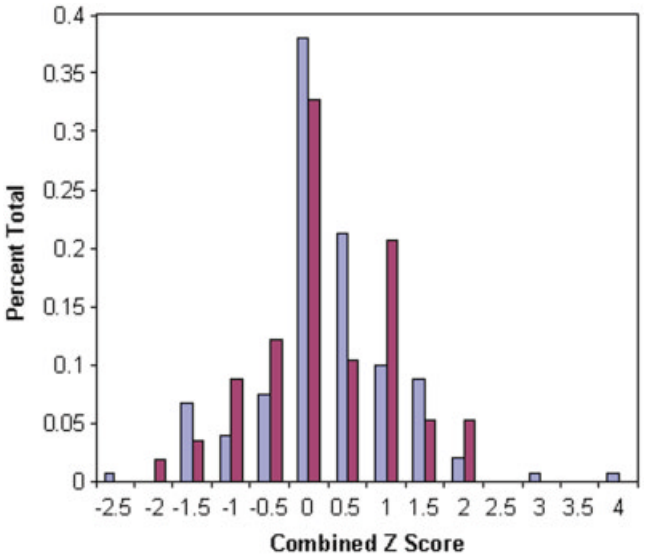

Fig. 7. Score distribution of manual predictors and automatic servers. A histogram depicting the percentage of the total number of predictions scored (*y* axis) for manual predictors (lavender) and automatic servers (maroon) falling within the indicated *z* score bin ranges (*x* axis). *z* scores represent the average scores of all measures calculated for best model predictions.

Prediction Ranks: Categories and Highlights

When summing the scores of first model predictions of all fold recognition domains, our assessment and evaluation of ranking significance suggested that the methods of three groups (Baker, Ginalski, and Rychlewski) outperformed the methods of all other groups. However, rankings changed when we used the best model predictions for each target or when we scored the predictions by averaging. Notably, two groups (group 425 Venclovas and group 448 Murzin) who did not score well by the summation approach emerge in the top 20 when the mean approach was used. Each of these groups predicted fewer target domains than the remaining groups (Venclovas with 7 domains and Murzin with 22 domains, out of 46 possible domains). Other than these exceptions, the rankings produced by each method followed the same general trends as most of the top performing groups predicted all target domains.

Table II includes overall scores, ranks, and bootstrap percentages for all best model fold recognition domain predictions of the top 20 groups. Considering best model

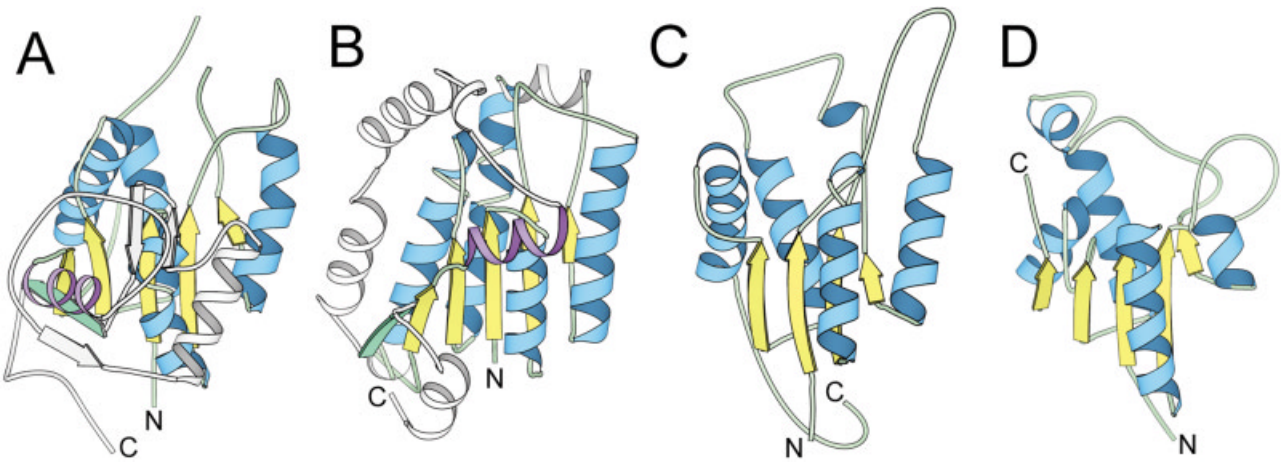

Fig. 8. Partial fold recognition. Target T0173 *M. tuberculosis* mycothiol deacetylase (**A**) experimental model, (**B**) group 349 Shortle first prediction model, (**C**) group 2 Baker first prediction model, and (**D**) group 12 Xu first prediction model. The secondary structural elements belonging to the N-terminal Rossmann-like fold are colored blue (α-helices), yellow (β-strands), and pale green (coil). Secondary structural elements C-terminal of the Rossmann-like fold are colored white (not predicted), purple (correct prediction), and green (prediction in incorrect orientation).

predictions, the maximum score was still achieved by group 2 (Baker, 3.58), followed closely by that of group 10 (Skolnick, 2.69). Figure 6 highlights difficulties of these two top scoring groups (Baker and Skolnick) with picking models. Group 10's (Skolnick) prediction model 3 [Fig. 6(A), green] and group 2's (Baker) prediction model 4 [Fig. 6(C), green] superimposed nicely with experimental target T0170, whereas their first model predictions superimposed poorly [Figs. 6(B) and (D), respectively]. It is interesting that each of these pairs of predictions corresponded to mirror images, where the group assigned the incorrect handed structure as the first model.

The top 20 groups summarized in Table II included three automatic servers. The first model predictions of group 29 (Robetta, 2.02) ranked high (finish ranked 4 in 33% of bootstraps), whereas the best model predictions of group 45 (Pmodel3, 1.87) and group 40 (Pmodel, 1.65) ranked 3 (38% of bootstraps) and 7 (14% of bootstraps), respectively. Although the average of the overall score distribution (first models, combined measures) for automatic servers (−0.28) was lower than that for manual groups (0.11), the performance of the top automatic server first model predictions (group 29 Robetta) approached those of the best manual predictors, and the performances of the top automatic servers best model predictions (group 45 Pmodel3, group 29 Robetta, and group 40 Pmodel) approached those of the best manual predictors. Figure 7 illustrates the score distribution of manual predictors and automatic servers on best model predictions.

The fold recognition category included a variety of domains encompassing a wide range of difficulty levels. Many of the top performing fold recognition groups also performed well in the comparative modeling and the new fold categories. We could emphasize this concept by splitting the fold recognition domains into homologues (30 domains) and analogs (16 domains) of known folds. Table III summarizes the resulting scores. Top achieving groups for fold recognition homologues (group 453 Ginalski, group 6 Rychlewski, and group 517 Bujnicki, Table III) performed well in the comparative modeling assessment,[25] and top achieving groups for fold recognition analogs (group 2 Baker, group 349 Shortle, and group 1 Karplus, Table III) performed well in the new fold assessment.[28] For fold recognition homologues, the score of one additional group (group 96 Bates) approached that of some of the best performers. For fold recognition analogs, group 2 (Baker) significantly outscored (5.09) all other methods, whereas the automatic server (group 29 Robetta) developed by the Baker laboratory ranked an impressive 3 (best models) and 5 (first models).

Because of improved sequence homology detection methods and increasing numbers of available sequences in databases, a significant overlap existed between the CASP5 comparative modeling (defined as any domain with detectable sequence similarity to a known fold) and fold recognition (defined by a natural boundary between domain clusters in Fig. 1). To minimize this overlap, we defined a subcategory of FR domains that included structures displaying obvious structural similarity in the absence of significant sequence similarity (25 domains). The corresponding predictions were evaluated by separating structural and alignment quality measures (Table IV). Although the same groups tended to perform significantly better than others (bold, Table IV) using structural measures (group 453 Ginalski, group 2 Baker, and group 6 Rychlewski) or alignment measures (group 2 Baker and group 453 Ginalski), one additional group (group 110 Honig) could be distinguished by their overall structural quality of predictions in this subcategory.

Highlights and Pitfalls of CASP5 Predictions

To exemplify the progress made in CASP5 and to identify potential focuses of future efforts in structure prediction, we highlight the predictions of several interesting fold recognition targets. Mycothiol deacetylase from *M. tuberculosis* (T0173) represents one of the most difficult examples. This single-domain protein, classified as FR(A)/NF, contains a novel Rossmann-like α/β fold with an overall topology similar to that of SAM-dependent methyltransferase but with a distinct sheet curvature. Many top scoring groups correctly predicted various aspects of the Rossmann-like N-terminus. However, no group correctly represented the overall fold topology. In fact, the closest prediction (group 349 Shortle, first model) to the overall fold topology in terms of coverage correctly extended the Rossmann-like fold by only one helix [purple, Fig. 8(A)] and incorrectly extended the sheet with a parallel β-strand (green, Fig. 8). This prediction ranked 7 among all first models, however, because it missed the distinct sheet curvature present in the experimental structure.

The first model prediction of group 2 (Baker) achieved the top score for this difficult target with all measures [Fig. 8(B)]. The Baker model correctly assigned four of five of the Rossmann-like β-strands with accurate sheet curvature, relatively good alignment, and correct stacking of α-helices. Model 1 submitted by group 12 (Xu) represented another noteworthy prediction for this target (rank 2 for combined measures). This prediction included all of the Rossmann-like β-strands in a curved sheet [Fig. 8(C)]. However, the lengths and stacking of the α-helices differed slightly, as indicated by a lower structural measure ranking for this individual target (rank 3 among first models and rank 4 among best models).

In CASP5, the correct prediction of multiple-domain proteins in general remained a challenging task. Some fold recognition targets contained rather difficult domain organizations, including swaps of secondary structural elements between two domains (T0148) or between two monomers (T0192 and T0193) and discontinuous boundaries in terms of primary sequence (T0146, T0148, T0159, T0168, T0172, T0174, T0186, T0187, and T0191). To evaluate the performance of the prediction community on such proteins, we compared the average GDT_TS scores attained on single-domain targets with those attained on multiple-domain targets in terms of target difficulty. Briefly, a trend line through the target difficulty data [Fig. 1(A)] represents the predicted average performance for each difficulty level. The relative performance of each

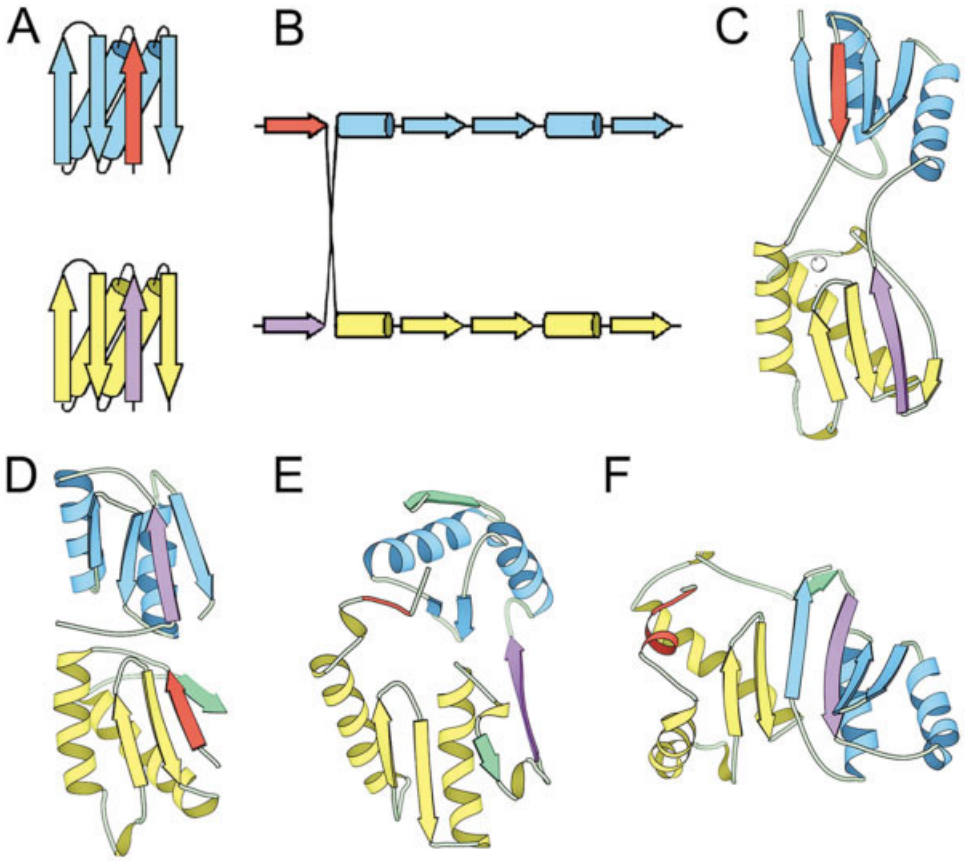

Fig. 9. Difficult domain organization. Secondary structural elements making up the tandem repeat of ferredoxin-like folds found in Target 148 are colored blue (first domain) and yellow (second domain), with the strand swap highlighted in red and purple. Secondary structural β-strands (arrows) and α-helices (cylinders) of the target structure are represented as (**A**) cartoon and (**B**) linear diagrams. Secondary structural elements of (**C**) the T0148 experimental model structure, (**D**) group 10 Skolnick first model fragment 1 and 2 predictions, (**E**) group 29 Robetta first model prediction, and (**F**) group 2 Baker first model prediction. Secondary structural elements of the target structure are colored according to the structural diagrams. Secondary structural elements of the prediction models are colored according to the corresponding residues in the target structure, with incorrectly placed (nonswapped) elements highlighted in green.

target with respect to this trend line was calculated by residuals (i.e., the difference between the calculated average performance and the actual performance), with positive values representing better than expected results and negative values representing worse than expected results [Fig. 1(B)]. We then averaged single-domain target residuals (0.194) and multiple-domain target residuals (−0.151) separately [Fig. 1(B), trend lines]. These results indicate a better performance of the CASP5 community as a whole on predicting single-domain targets.

To demonstrate the difficulties of complex domain organizations, we discuss one interesting target (T0148) that contains a tandem repeat of ferredoxin-like fold domains [each classified as FR(A)] with swapped N-terminal β-strands [Figs. 9(A)–(C)]. Although no group correctly predicted the strand swap, several top performing groups predicted the ferredoxin-like fold duplication [group 10 Skolnick, Fig. 9(D); group 2 Baker, group 6 Rychlewski, group 450 Labesse, group 28 Shi, group 427 Fischer, and group 110 Honig, not shown]. The server Robetta (group 29) produced a remarkable first model prediction for this target. Although the model contained incomplete sheets, it correctly assigned the discontinuous boundaries of the two domains [Fig. 9(E)], and outperformed the manual prediction from the same group [group 2 Baker, Fig. 9(F)]. This result highlights a potential use of such fragment-based prediction methods in domain parsing.

Several predictions more closely resembled the target structure than any available template. The first model prediction of the FR(H) target T0134_2 by group 453 (Ginalski) represented one such accomplishment (Fig. 10). Ginalski combined two existing PDB template structures (1qts and 1e42) to produce a model structure with a GDT_TS score (79.01) slightly better than that of the closest template (1qts, 78.19). Similarly, group 28 Shi achieved a slightly higher GDT_TS score (63.96) than the closest available PDB template 1hjr (63.89) with a second model prediction of the target T0157. In view of the numerous insertions/deletions and the relatively low-sequence identities between each of these targets and their

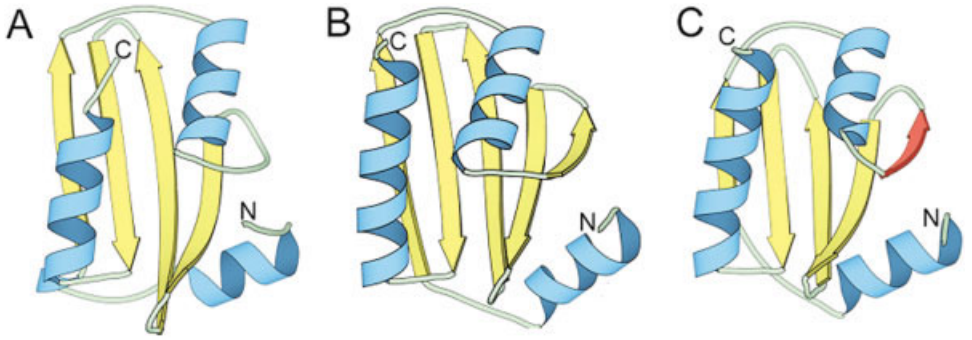

Fig. 10. Prediction beats available templates. Structural models are depicted for (**A**) the closest available template (1qts) to target domain T0134_2, (**B**) the experimental domain T0134_2, and (**C**) the best prediction (group 453 Ginalski first model) for T0134_2. Helices and strands common to all structures are colored blue and yellow, respectively. Compared with the corresponding experimental model structure (**B**), secondary structural elements with incorrectly aligned residues are colored red.

templates (12.7% for T0134_2 and 1qts; 15.4% for T0157 and 1hjr), both of these high scoring predictions achieved remarkable alignment quality. For example, the Ginalski model for T0134_2 correctly aligned all but one of the secondary structural elements [black, Fig. 10(C)]. Considering this alignment shift, the higher GDT_TS score becomes more significant in a structural sense. In fact, the target domain superimposed with the Ginalski model (RMSD 1.6 over 105 residues) better than it superimposed with either template used to generate the model (1qts: RMSD 1.8 over 100 residues and 1e42: RMSD 1.8 over 104 residues).

The CASP5 fold recognition targets included an example of a singleton sequence (T0174) that did not find any similar sequences in public databases using PSI-BLAST procedures. Such an isolated sequence represents a particular challenge for fold recognition methods. The T0174 experimental structure belongs to the two-domain SCOP superfamily of GHMP kinases. Remarkably, many groups (11 automatic servers and 10 manual predictors) correctly identified the GHMP kinase fold as a template for this protein (reported as parent 1kvk, 1kkh, 1fwk, 1fi4, 1k47, and 1h73). Of these groups, two automatic servers (group 45 Pmodel3 and its input server group 39 Pcons3) and several manual predictors produced first models with reasonable overall folds (group 373 Brooks, group 437 Ho-Kai-Ming, group 96 Bates, and group 10 Skolnick; group 453 Ginalski and group 6 Rychlewski reported N/A as parent template but identified the correct fold). Considering best models, more automatic servers found the correct fold and obtained satisfactory scores (group 362 PSPT, group 40 Pmodel, group 38 Pcons2, group 14 FUGUE2, group 226 FUGUE3, and group 221 INBGU) than did manual predictors (group 47 Gibrat, group 464 Catherinot, and group 8 Royyuru).

Although many predictors identified a correct template for the singleton target T0174, the resulting structural models diverged significantly from the target structure. GDT_TS scores for the best domain predictions (group 45 Pmodel3, 26.015 for T0174_1; and group 437 Ho-Kai-Ming, 40.161 for T0174_2) did not approach

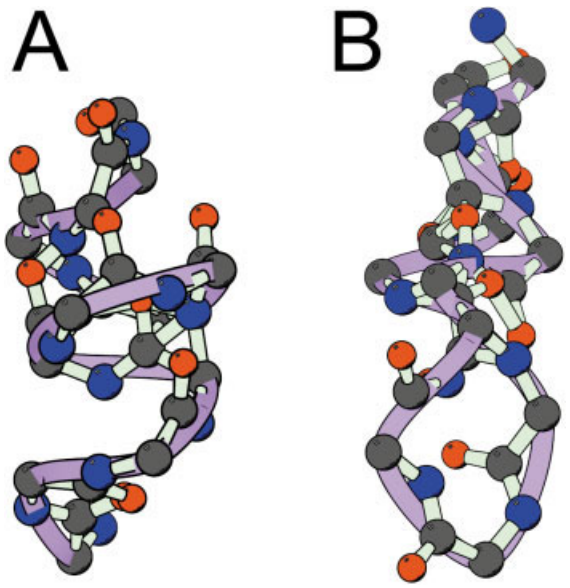

Fig. 11. Physically unrealistic models. Ball-and-stick representations of backbone coordinates from one best model prediction for Target T0174 illustrate (**A**) two turns of a helix-like secondary structure from residue 350 to residue 360 and (**B**) two overlapping strand-like secondary structures connected by a turn from residues 176 to 186. Atoms are colored according to type: red (oxygen), blue (nitrogen), and gray (carbon). Bonds are drawn between atoms <1.6 Å apart (pale green), and the backbone trace is illustrated in purple.

those calculated for a manual structural alignment with the closest template (1kvk, 55.492 for T0174_1 and 68.614 T0174_2). For most of the predictions with correct parent templates, these poor scores reflected both incomplete alignment coverage and poor alignment quality. For example, the coordinates of one automatic server prediction (group 189 SAM-T02-server best model) only extended over one domain of the target structure, whereas another group (group 537 Asogawa) misaligned the entire target sequence with that of the parent template (1kvk). Although the structure of this misaligned prediction looked similar to that of the top

predictions, the poor alignment was reflected in a low overall alignment quality score (z score of -0.34).

In addition to poor alignment quality, many top predictions for Target T0174 displayed poor local structural quality. For example, one ab initio method produced a physically unrealistic model that generated one of the top overall scores for this target (average z score for both domains 2.035) (Fig. 11). Figure 11 illustrates examples of two secondary structure regions in this model where backbone coordinates essentially overlap (fall within a bond distance of 1.6 Å). Despite the presence of these poor structural regions, the model correctly represented the β-strand with respect to the template structure (1kvk) for the second domain (T0174_2). Such a result underscores a generalized need for applying some form of model refinement to structures produced by a number of fold recognition and ab initio-type methods (see also new fold assessment article[28]).

DISCUSSION

This report concentrates on evaluating the top scoring groups in the CASP5 fold recognition category to establish the current state of the art in protein structure prediction methods. Despite the obvious advantages of such an assessment, the evaluation process falls short in identifying promising new methods that may ultimately drive future progress in the field. One of the original objectives of CASP was to identify the best methods in the three different categories: comparative modeling, fold recognition, and ab initio protein structure prediction. Today, this goal is masked by a growing trend to combine various existing prediction tools in unique ways. Both manual groups and automatic servers that exploit such combinations surpass the independent "threading" methods that once dominated the fold recognition field.

In many ways, the CASP assessment process itself has driven the tendency of predictors to merge existing prediction techniques. Pressures to predict each and every target force groups to use a combination of the best available techniques for each prediction category. Fortunately, such pressures did not distract from developing and improving methods in CASP5. The group 29 automatic server Robetta provides an excellent example of this point. Robetta combined results of group 38 Pcons2, an automatic server that performs well on comparative modeling targets, with results of the ab initio method Rosetta developed by the Baker group. This combination produced an outstanding prediction for target T0186, a three-domain protein classified as CM and CM/FR(H) for two domains and FR/NF for a third, inserted domain. Observations in previous CASP assessments that different groups performed better on different targets also reinforced the notion that sharing techniques could lead to better performance. The 3D JURY method developed and used by group 6 Rychlewski and by group 453 Ginalski took full advantage of this idea.[29] By using a detailed scoring scheme, 3D-Jury used input models produced by a large set of publicly available fold recognition servers and picked those that had the most abundant high scores as predictions.

The results of the fold recognition assessment suggest that two general types of methods performed well in protein structure prediction: template-based methods and fragment assembly-based methods. For fold recognition homologues, groups using methods that combined models built from existing templates with some type of model refinement performed well. The procedure used by group 453 Ginalski, one of the top scoring groups in both fold recognition and comparative modeling, provides an excellent example of such a method. Ginalski started with templates provided by comparative modeling and fold recognition servers, improved template alignments manually by using various criteria, built homology models from these alignments, and improved the resulting models with a combination of available programs and manual inspection. As discussed, this procedure produced a prediction closer to the experimental target than the closest template (Fig. 10).

The second general type of method has been categorized as ab initio protein structure prediction. This method type generally concentrates on local structural properties and performs well on fold recognition analogs and new folds. The Rosetta program developed by group 2 (Baker) and the TOUCHSTONE folding algorithm developed by group 10 (Skolnick) are both examples of such methods. Rosetta identified small fragments from a library generated from existing structures,[30,31] whereas the ab initio folding segment of TOUCHSTONE identified conserved contacts from multiple weakly significant template fragments found by threading.[32]

With the overall performance of fold recognition groups being easily divided into two main classifications (methods that work on homologues and methods that work on analogs) and with the increasing depletion of fold recognition domains into the comparative modeling group, perhaps a change in CASP assessment categories is required. Under the current definition, significant overlaps exist between assessment categories, especially between those of fold recognition and comparative modeling. Although this overlap provides a nice check for assessment methods, it detracts from the very different aspects of model comparisons that need to be made for each category. For example, the local structural details of side-chains and loops cannot be assessed in most of the target domains currently defined as comparative models due to significant structural divergence. Although such a localized comparison is essential for the methods development of comparative modeling groups, the task becomes overwhelmed by the huge number of target domains assigned to the category that must be addressed on a more global level. Therefore, we suggest that targets be categorized on the basis of the methods required for assessment, rather than the outdated methods used for predictions. This type of classification scheme fits more naturally within the outline of CASP, because assessors have no knowledge of methods when establishing ranks.

In addition to the promising predictions of the top scoring groups outlined in this report, the prediction community as a whole achieved higher overall average

scores (GDT_TS) in CASP5 than in CASP4 (see CASP5 progress article[33]). However, the question remains about whether this better performance actually reflects significant advances in prediction technology. Groups participating in CASP5 have larger databases of sequence and structure information available, in addition to a number of publicly accessible fold recognition servers. One of the most promising results of the CASP5 assessment of fold recognition concerns the performance of automatic servers. Predictions by the best performing servers approach those of the top scoring manual groups. In fact, group 29 Robetta performed among the top groups (with a statistically significant rank 3) for the subset of "real" fold recognition domains using alignment measures on first models, whereas group 45 Pmodel3 and group 40 Pmodel performed consistently among the top groups using best models.

This assessment identified the top scoring groups for all fold recognition domains and for various subsets of domains, including homologues and analogs, using various structural and sequence alignment criteria. General trends in average scores attained by all groups on various targets were described. A huge amount of data produced in the course of our analysis of fold recognition predictions exists that cannot be contained within this report. Further analysis of this data could provide additional insights into general trends of protein structure prediction and into specific strategies of individual prediction groups. Interested readers can access the data generated for this assessment in the files at http://predictioncenter.llnl.gov/casp5/fr/.

ACKNOWLEDGMENTS

We thank the CASP 5 organizers, John Moult and Tim Hubbard, for asking us to be a part of the CASP experience. We greatly appreciate input and discussions from many CASP participants: Alexey Murzin, Rob Russell, Patrick Aloy, Anna Tramontano, Veronica Morea, Krzysztof Fidelis, Česlovas Venclovas, and Adam Zemla. Additional thanks goes to the crystallographers and NMR spectroscopists who provided structure data for CASP 5 and to the predictors who provided a tremendous amount of work (and fun) for us.

REFERENCES

1. Kinch LN, Qi Y, Hubbard T, Grishin NV. CASP5 target classification. Proteins 2003;Suppl. 6:340–351.
2. Hubbard TJ. RMS/coverage graphs: a qualitative method for comparing three-dimensional protein structure predictions. Proteins 1999;Suppl 3:15–21.
3. Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins 1999;34 2:220–223.
4. Zemla A, Venclovas, Moult J, Fidelis K. Processing and evaluation of predictions in CASP4. Proteins 2001;Suppl 5:13–21.
5. Zemla A. LGA program: a method for finding 3-D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
6. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins 1999;Suppl 3:22–29.
7. Holm L, Park J. DaliLite workbench for protein structure comparison. Bioinformatics 2000;16:566–567.
8. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
9. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
10. Ortiz AR, Strauss CE, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 2002;11 11:2606–2621.
11. Sippl MJ, Lackner P, Domingues FS, Prlic A, Malik R, Andreeva A, Wiederstein M. Assessment of the CASP4 fold recognition category. Proteins 2001;Suppl:55–67.
12. Levitt M. Competitive assessment of protein fold recognition and alignment accuracy. Proteins 1997;Suppl 1:92–104.
13. Murzin AG. Structure classification-based assessment of CASP3 predictions for the fold recognition targets. Proteins 1999;Suppl 3:88–103.
14. Lemer CM, Rooman MJ, Wodak SJ. Protein structure prediction by threading methods: evaluation of current techniques. Proteins 1995;23:337–355.
15. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
16. Altschul SF, Koonin EV. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. Trends Biochem Sci 1998;23:444–447.
17. Lackner P, Koppensteiner WA, Sippl MJ, Domingues FS. ProSup: a refined tool for protein structure alignment. Protein Eng 2000;13:745–752.
18. Grishin VN, Grishin NV. Euclidian space and grouping of biological objects. Bioinformatics 2002;18:1523–1534.
19. Koehl P. Protein structure similarities. Curr Opin Struct Biol 2001;11:348–353.
20. Zemla A. Proteins 2003;Suppl 6.
21. Rost B, Sander C, Schneider R. Redefining the goals of protein secondary structure prediction. J Mol Biol 1994;235:13–26.
22. Sauder JM, Arthur JW, Dunbrack RL Jr. Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins 2000;40:6–22.
23. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. LiveBench-2: large-scale automated evaluation of protein structure prediction servers. Proteins 2001;Suppl 5:184–191.
24. Tramontano A, Leplae R, Morea V. Analysis and assessment of comparative modeling predictions in CASP4. Proteins 2001;Suppl 5:22–38.
25. Tramontano A, Morea V. Assessment of homology based predictions in CASP5. Proteins 2003;Suppl 6:352–368.
26. Marti-Renom MA, Madhusudhan MS, Fiser A, Rost B, Sali A. Reliability of assessment of protein structure prediction methods. Structure (Camb) 2002;10 3:435–440.
27. Kendall MG, Stuart A, Ord JK. Kendall's advanced theory of statistics. New York: Oxford University Press; 1987. v. p. 77–80, 365–368.
28. Russell RB, Aloy P. Predictions without templates: new folds, secondary structure and contacts in CASP5. Proteins 2003;Suppl.
29. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.
30. 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.
31. Simons KT, Bonneau R, Ruczinski II, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 1999;37:171–176.
32. Kihara D, Lu H, Kolinski A, Skolnick J. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. Proc Natl Acad Sci USA 2001;98:10125–10130.
33. CASP5 Progress. Proteins 1993.
34. Esnouf RM. An extensively modified version of MolScript that includes greatly enhanced coloring capabilities. J Mol Graph Model 1997;15:132–134, 112–113.