PROTEINS: Structure, Function, and Genetics 53:436–456 (2003)

Predictions Without Templates: New Folds, Secondary Structure, and Contacts in CASP5

Patrick Aloy, Alexander Stark, Caroline Hadley, and Robert B. Russell*
EMBL, Heidelberg, Germany

ABSTRACT **We present the assessment of CASP5 predictions in the new fold category. For coordinate predictions, we considered five targets with new folds and eight lying on the fold recognition borderline. We performed detailed visual and numerical comparisons between predicted and experimental structures to assess prediction accuracy. The two procedures largely agreed, but the visual inspection identified instances where metrics, such as GDT_TS, ranked what we considered incorrect predictions highly. We found the quality of the best predictions to be very good: for nearly every target at least one group predicted a structure close to the correct one. However, selection of the best of five models is still problematic. The group of David Baker once again proved to be best overall, with many individual highlights. However, high quality and consistency were also seen from others, suggesting that the community is moving toward general procedures to predict accurate structures for proteins showing no resemblance to anything seen before. Predictions for secondary structure showed at best limited progress since CASP4. The number of targets is probably too small to spot differences in performance between methods, suggesting that such predictions might be better evaluated with schemes involving more proteins. For contact predictions, accuracies are still low, although there were several instances of accurate and useful contacts predicted de novo, and new approaches hint at future progress. Proteins 2003;53:436–456.**

Key words: de novo/ab initio structure prediction; secondary structure; residue–residue contacts; CASP5 assessment

INTRODUCTION

The quest for a general solution to the protein-folding problem has lured scientists from many disciplines for nearly five decades. Despite many overstated claims, progress toward this goal was slow until the past decade. The CASP new fold category (formerly ab initio) has played a key role in charting progress in the ability to predict protein structures without a suitable starting template. CASP experiments have forced predictors to exercise caution when making claims of success that might not stand up under the scrutiny of such blind trials; as a result, they have removed much of the soothsaying reputation that had existed previously.

The methods evaluated during CASPs 1 and 2[1,2] in the new fold category gave accurate predictions only for small folds, fragments, or supersecondary structures, and methods typically showed little consistency. CASPs 3 and 4[3,4] saw the emergence of a new class of methods that predicted structures based on the assembly of fragments built up after first finding parts of other structures compatible with parts of the target sequence (e.g., Refs. 5 and 6). These predictions could be very accurate, producing structures of sufficient quality to be used in experimental design or the discovery of unexpected relationships (e.g., Ref. 7). It was in this context that we evaluated the new fold category predictions in CASP5. Knowledge of these previous successes meant that we tended to be rather ruthless. We considered predictions successful only when they were near to the correct fold and did not give too much credit for pieces of local structure or supersecondary structures that might arguably be just a logical extension of a good secondary structure prediction.

We also evaluated predictions of secondary structure and interresidue contacts. Here our main aim was to compare the results to CASP4 in the simplest possible way in order to chart general progress in the field and to spot new innovations easily.

MATERIALS AND METHODS

Classification of Targets

Table I lists the targets considered during the analysis of coordinate predictions. We ignored all targets showing unambiguous similarity to previously known structures [i.e., comparative modeling (CM) targets and clear fold recognition (FR) homologues and analogs]. This procedure left a total of 13 targets, of which 5 were defined as true new folds (NF) and 8 showed partial similarity to known structures and were defined as on the borderline with fold recognition (NF/FR). Target T0131 was ignored because of problems with the structure determination; T0139 was omitted because of publication before the deadline (indeed at least two groups used the article[8] to build a model of the structure; Alexei Murzin and Kevin Karplus, personal communication). Eight of the targets (2 NF and 6 NF/FR) were domains from larger proteins. All domains apart from the three in T0146 could, in principle, have been

*Correspondence to: Robert B. Russell, EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany. E-mail: russell@embl.de

Received 23 January 2003; Accepted 15 May 2003

TABLE I. The 13 Targets Considered During Coordinate Predictions

Target	PDB code	Name	Species	Class	Other domains
Five new fold targets (NF)					
T0129	1izm	HI0817	*H. influenzae*	all-α	
T0149_2 (203–318)	1nij	YjiA	*E. coli*	α+β	1 FR(H)
T0161	1mw5	HI1480	*H. influenzae*	α+β	
T0162_3 (114–281)	1izn	F-actin capping protein α1	Chicken	α+β	1 & 2 FR(A)
T0181	1nyn	YH07/YHR087w	*S. cerevisiae*	α+β	
Eight fold recognition borderline targets (NF/FR)					
T0146_1 (1–24, 114–196)	1nrk	YgfZ	*E. coli*	α+β	
T0146_2 (25–113)	1nrk	YgfZ	*E. coli*	α+β	
T0146_3 (244–299)	1nrk	YgfZ	*E.coli*	α+β	
T0170	1h40	HYPA/FBP11 FF domain	Human	α	
T0172_2 (116–216)	1m6y	MRAW	*T. maritima*	α	1 CM/FR(H)
T0173	-	Mycothiol deacetylase	*M. tuberculosis*	α/β	
T0186_3 (257–292)	1o12	TM0814	*T. maritima*	all-β	1 CM, 2 FR(H)
T0187_1 (4–22, 250–417)	1o0u	TM1585	*T. maritima*	α/β	1 FR(H)

identified by first finding other domains by sequence comparison or FR methods (see Table I for details). In the sections that follow, we refer to targets by their number (e.g., T0149) and append this with the domain number when necessary (e.g., T0149_2).

Coordinate Predictions

We compared predicted and target structures by using several measures, including RMSD,[9] LCS,[10] SOV,[11] and GDT_TS.[12] Our initial analyses convinced us that GDT_TS gave the best agreement with visual inspections of the models, and it thus became central to all evaluations.

We attempted several schemes to assess the overall performance of groups using GDT_TS. The first was based on a *z* score (number of standard deviations above the mean), where the score for overall group performance was the sum of the best *z* score for each target. We did not add negative values to this total because we did not consider it fair to penalize groups for wrong predictions; moreover, we did not believe differences between negative values were meaningful.

Two other ranking schemes were derived from individual per target GDT_TS scores. The first of these was a simple score for the top 50% of predictions where the raw GDT_TS scores were normalized to 100 for each target and summed over all targets. The second was akin to that used in secondary structure prediction here (see below) and in CASP 4[4]: models were awarded points according to the percentile rank of GDT_TS: values ≥ 50, ≥ 60, ≥ 70, ≥ 80, ≥ 90 earned +1, +1, +1, +1, and +2 points, respectively.

We also performed a detailed visual assessment of the predictions. Models sorted according to GDT_TS were inspected (by RBR) using molecular graphics,[13] and superimpositions were generated by either the CASP5 Web site (http://predictioncenter.llnl.gov/casp5; sequence dependent or independent) or when necessary by structure-based alignment (STAMP[14]). Models scored 2 points if the overall fold was largely correct (labeled "excellent"), and the prediction could arguably be useful in the design of biological experiments. Models that predicted difficult features or were deemed to be part of the way toward the correct fold scored 1 ("good"), and all others scored 0. Inspection of models stopped when it became clear that no further points would be awarded (typically 10–20 scoreless predictions). The number of models inspected varied from 20 to 124, the total roughly correlating inversely with the difficulty of the target. Approximately 1000 models were inspected in total.

As in the CASP4 assessment, we gave equal weight to all models when determining the overall ranks, but we tested whether groups were able, as suggested at the outset of the experiment, to select a best model as their first.

Secondary Structure Predictions

As in CASP4,[4] we ranked the group performances with use of SOV percentiles. A group was awarded 1 point if its SOV score lay above the 50th percentile (the median), plus 1 point if it was above the 60th, 70th, or 80th percentiles and 2 points if above the 90th (i.e., the group performed better than 90% of the groups). We assigned a maximum of 6 points to every prediction. We then normalized these scores from 0 to 100 for each target. We gave the total group score as the average of these values, either over the total number of targets or over the number of targets predicted by the group. We also used SOV score percentiles to define the intrinsic prediction difficulty of each target. We considered a target as easy if the 75th percentile score was $\geq$85 (i.e., 75% of the groups scored 85 or better) or hard if the 75th was $\leq$68 and the 90th was $\leq$75. Otherwise, it was considered of medium difficulty.

Residue–Residue Contact Predictions

For the residue–residue contact predictions, we defined measures of prediction accuracy (fraction of predicted contacts that are true positives) and coverage (fraction of observed contacts predicted) as:

$$\text{Accuracy} = \text{Tp}/\text{N}_{\text{pred}};\ \text{N}_{\text{pred}} = (\text{Tp} + \text{Fp})$$

$$\text{Coverage} = \text{Tp}/\text{N}_{\text{obs}};\ \text{N}_{\text{obs}} = (\text{Tp} + \text{Fn})$$

where Tp = true positives; N_{pred} = total number of predicted contacts; N_{obs} = total number of observed con-

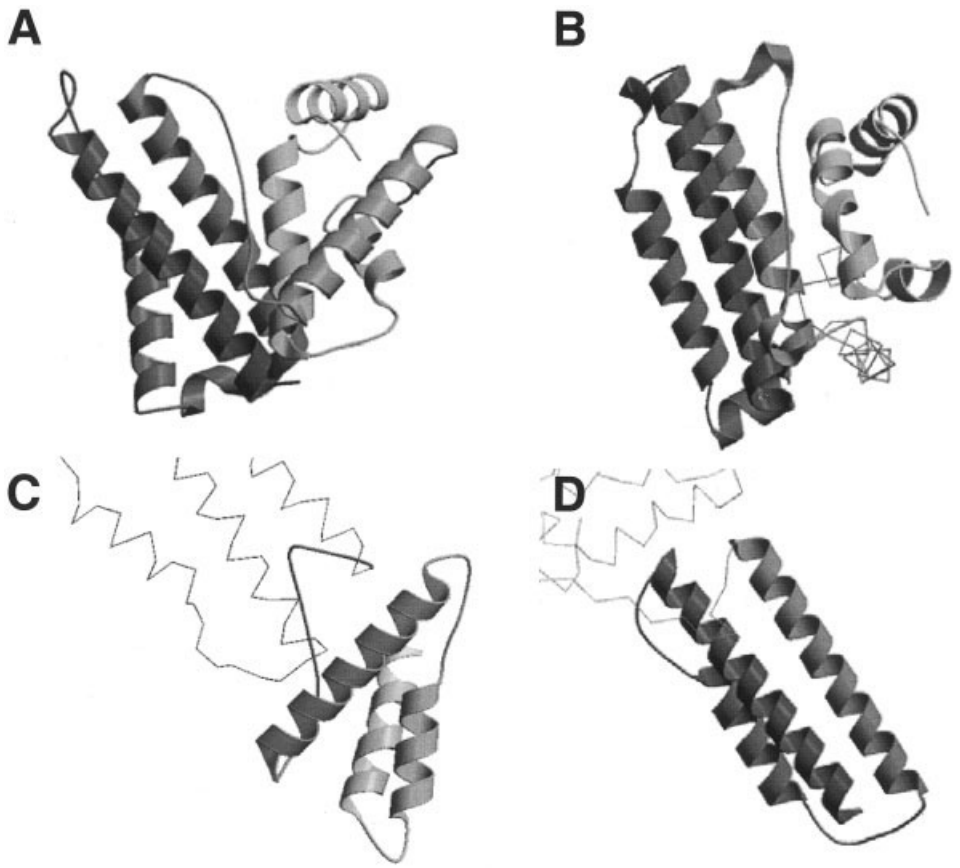

Fig. 1. Predictions for T0129. Correct structure (**A**) and predictions from the groups of (**B**) Baker (002, model 4) and (**C** and **D**) Doniach (401, model 2). The first is an example of a correct overall prediction; the second one shows where the N-and C-terminal halves have been predicted separately but not put together correctly. Figures were made by using Molscript[33] and Render.[34] Structures are colored from light to dark gray as one moves from N-to C-terminus, and regions in the predicted structures not considered to be correct are shown in Cα trace.

tacts; Fn = false negatives (underprediction), Fp = false, positives (overprediction). A true positive is an observed contact correctly predicted; a false positive is a predicted contact that is not observed in the target (overprediction) and a false negative is an observed contact that has not been predicted (underprediction).

Additional data related to our assessment can be found on an accompanying Web site: http://www.russell.embl.de/casp5/.

RESULTS

Coordinate Predictions

For the 13 targets considered, there were a total of 4840 models, from 165 groups. Groups were roughly equally split regarding whether they submitted a single model or several: in 691 cases the groups submitted 5 models, and in 565 they submitted just one (with 94, 56, and 138 instances of 4, 3, and 2 models, respectively).

Results for Individual Targets

The sections that follow discuss results for each of the 13 targets individually and highlight successful predictions and interesting observations.

T0129 (NF): HI0817 from H. influenzae

This protein contains seven α-helices. The first four and the last three form separate bundles (N-and C-terminal bundles) separated by a long, essentially extended crossover loop between helices 4 and 5 [Fig. 1(A)]. An additional noteworthy feature is another extended loop segment connecting helices 6 and 7.

Inspection of the results showed this target to be difficult when considered as a whole, with just one group getting the overall fold right. The best predictions were models 4, 1, and 3 from the Baker group (002), which are all variants on a theme that is essentially correct [Fig. 1(B)] apart from minor alterations in the N-terminal bundle, where helices 3 and 4 are swapped. Both extended loops were predicted, but by far, the most impressive aspect of these predictions is the correct relative orientation of the two subdomains. Many other groups made good predictions of the N-and C-terminal bundles but did not get the relative orientation correct. Good examples of this come from the Jones-NewFold (068) and Doniach (401) groups [Figs. 1(C) and (D)], both of whom—like several other groups—predicted a largely correct structure for the extended linker region.

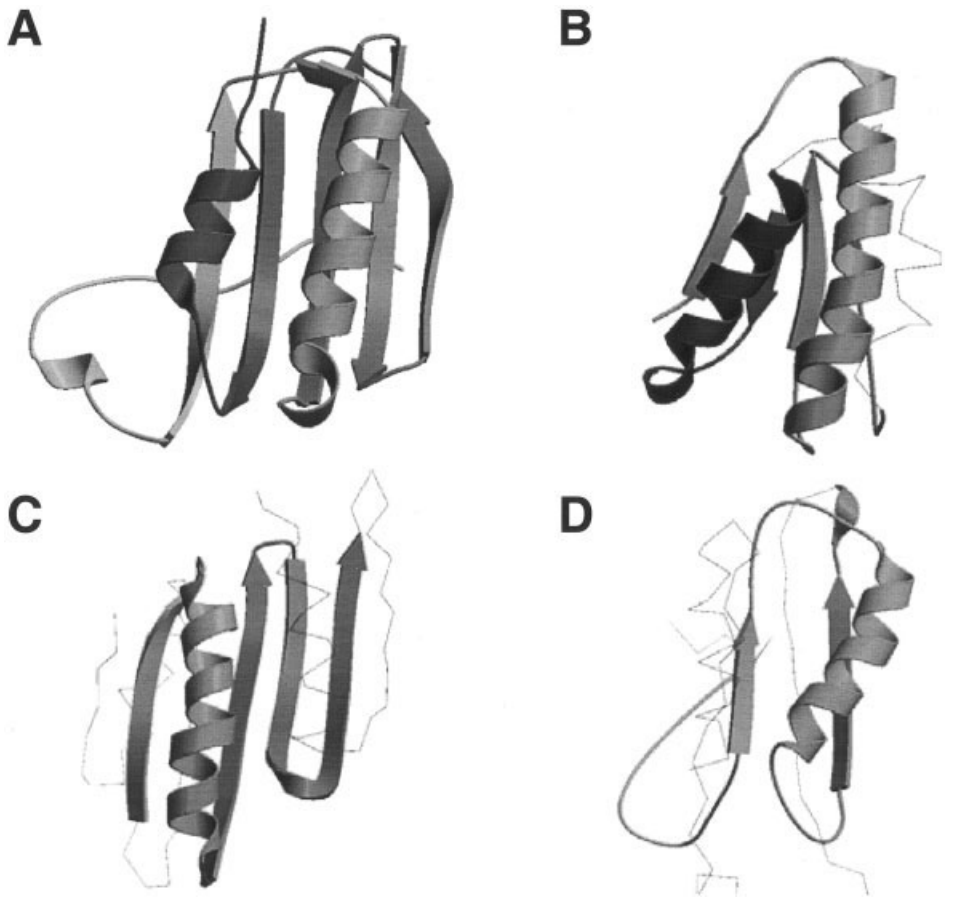

Fig. 2. Predictions for T0149_2. Correct structure (**A**) and predictions from the groups of (**B**) Shortle (349, model 2), (**C**) Scheraga (314, model 2), and (**D**) Baker (002, model 1). The first was ranked fourth in GDT_TS assessment but was considered the closest match in topology by visual inspection. Details are as for Figure 1.

T0149 (NF): yjiA from E. coli, *domain 2*

This α + β protein consists of a single mixed β-sheet with strand order 15234 where strands 3 and 5 are antiparallel to the others [Fig. 2(A)]. Two α-helices, one between strands 1 and 2 and another at the C-terminus, pack against one side of the sheet, and an extended, somewhat irregular N-terminal segment packs against the other.

The best prediction as judged by the visual assessment was model 2 from the Shortle group (349), where the only topological variation is the replacement of edge β-strand 4 by an α-helix [Fig. 2(B)]. Model 2 from Scheraga (314) and model 1 from Baker (002) were also "good" predictions (i.e., 1 of 2 points), although these had more serious topological variations [Figs. 2(C) and (D)]. The latter two models scored higher according to GDT_TS (ranks 1 and 2) than that from the Shortle group (rank 4; another Scheraga model was at rank 3). This was probably due to some distortions in the Shortle model relative to the known structure, and some long, well-fitting segments in the Scheraga and Baker models [e.g., the long N-terminal coil segment predicted by the Baker group; Fig. 2(D)].

T0161 (NF) HI1480 from H. influenzae

This protein consists of several α-helices in an irregular arrangement and a three-stranded, antiparallel β-sheet of strand order 123 [Fig. 3(A)]. Strands 2 and 3 are very long and play a key role in forming a homodimer. The dimeric structure, together with the fact that this protein was a singleton (i.e., no sequence homologues at all), made this target extremely difficult to predict. The best predictions were from the Baker (002) and Jones-NewFold (068) groups, who both managed to get the approximate trace of the chain correct, although with many fine details wrong [Figs. 3(B) and (C)].

T0162 (NF) CAZ1 from chicken, domain 3

This domain consists of a central five-stranded β-meander, with N-and C-terminal α-helices packing against one face; the strands are very long and the sheet is very flat, and the C-terminal helix of 37 residues is exceptionally long [Fig. 4(A)].

The Brooks group (373) had a good prediction of the overall structure based on a template (human zinc-alpha-2-glycoprotein: 1zag), giving a structure lacking only the C-terminal helix [Fig. 4(B)]. The I-sites/Bystroff server

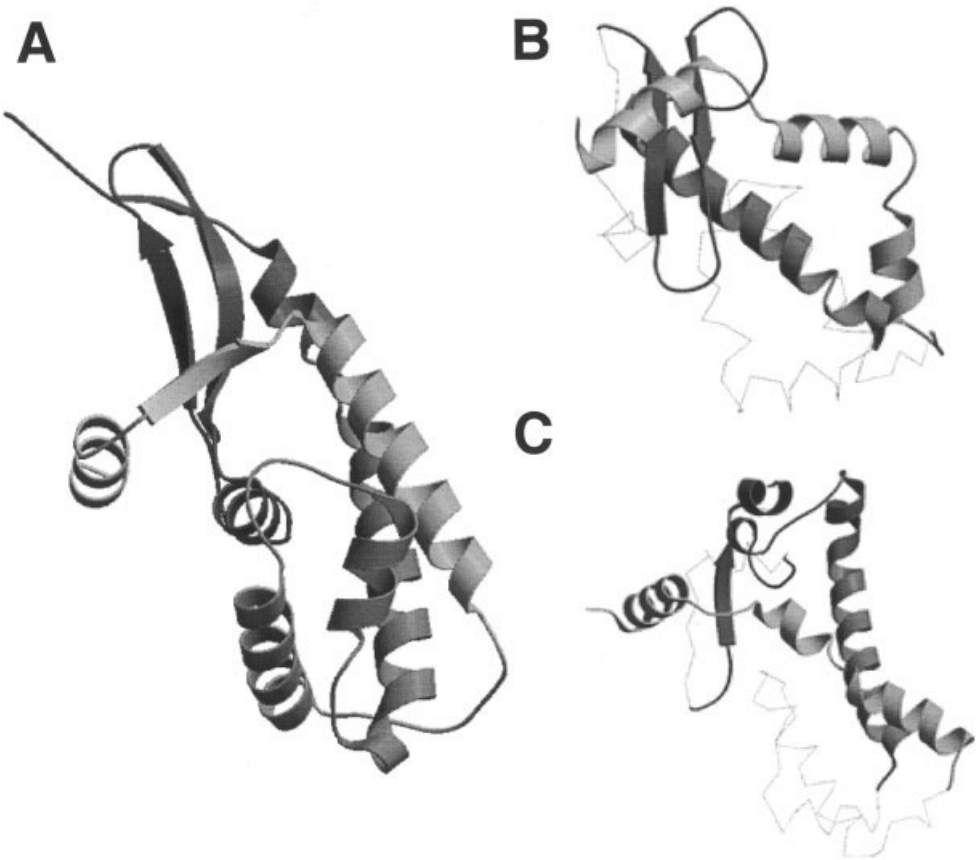

Fig. 3. Predictions for T0161. Correct structure (**A**) and predictions from the groups of (**B**) Baker (002, model 2), and (**C**) Jones (068, model 3). Details are as for Figure 1.

(132) predicted all but the last strand and helix correctly [Fig. 4(C)], and other groups predicted features of the structure but not the overall fold. For example, the SAM-T02-human group (001) predicted the structure correctly but in separate halves not forming a single sheet. Far down the GDT_TS ranked list the 3D-PSSM server (229) predicted a similarity to human ubiquitin-conjugating enzyme (1u9a), which has an identical topology, although the strands and helices are of very different lengths [Fig. 4(D)].

T0181 (NF) YH07 from S. cerevisiae

The N-terminus of this protein (residues 1–95) consists of a four-stranded β-meander (i.e., strand order 1234) with four α-helices packing against one side; two helices are inserted between strands 2 and 3, and the other two lie after strand 4 in the sequence [Fig. 5(A)]. The C-terminus (96–111) consists of something resembling a short, three-stranded β-meander, although main-chain hydrogen bonding does not define any strands here.

The best prediction came from the Skolnick-Kolinski group (010) who predicted the overall trace of the N-terminal half of the protein correctly, with a swap of strands 1 and 2 and the substitution of strands 3 and 4 by a helix/loop segment [Fig. 5(B)]. Other good predictions were made by SAM-T02-human [001; Fig. 5(C)] and Samudrala-NewFold [051; Fig. 5(D)] as well as the PROTINFO-AB server (140). The Jones-NewFold (068) and ORNL-PROSPECT (012) groups also predicted features well, the latter notably being one of a few that predicted the entire sheet topology correctly, despite differences in the rest of the structure.

T0146 (NF/FR) ygfZ from E. coli *domains 1, 2, and 3*

The three domains of this protein are all variations on an α + β theme and are probably descendants of a common ancestor[15] based around a structure resembling the ferredoxin-like fold. They were exceptionally hard targets: it was difficult, if not impossible, to identify the domains from sequence comparison or fold recognition. Domain 1 consisted of fragments from disparate parts of the polypeptide chain (i.e., domain 2 [Fig. 6(A)] was inserted into it), and domain 3 appeared to be a slightly decayed version of the other two, with the two helices replaced by unusual loop segments.

No group made correct predictions for domain 1 or 3, and indeed it was telling that the highest scoring (but incorrect) models were those based on templates (i.e., from FR methods), mostly similar to ferredoxin-like folds. However, for domain 2, the Shortle group (349) produced a prediction close to the correct topology [Fig. 6(B)].

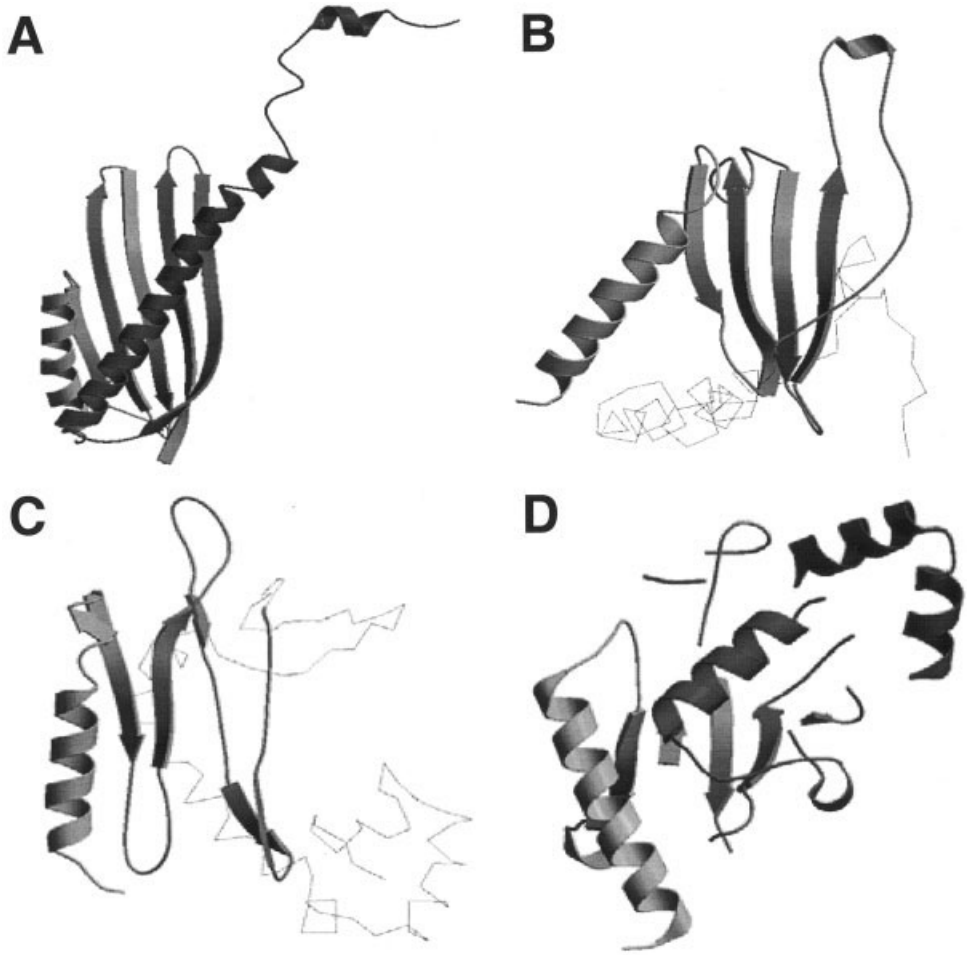

Fig. 4. Predictions for T0162_3. Correct structure (**A**) and predictions from the groups of (**B**) Brooks (373, model 4), (**C**) I-sites/Bystroff (132, model 1), and (**D**) 3D-PSSM (229, model 4). The last prediction is ranked 44th by GDT_TS but is the only one of these predictions that has the correct overall topology (although with many other differences in structural detail). The prediction in (B) was based on a template (1zag). Details are as for Figure 1.

T0170 (NF/FR) HYPA/FBP11 FF domain from human

This small protein consists of four helices arranged in a right-handed super-helix [Fig. 7(A)]. Apart from an unusual seven-residue, extended, slightly disordered segment at the N-terminus, the structure is quite regular. This was probably what made it one of the easiest targets, being predicted very accurately by the Baker [001; Fig. 7(B)], Skolnick-Kolinski (010), Samudrala-NewFold (051), and Wolynes-Schulten [294; Fig. 7(C)] groups and the PROTINFO-AB server (140) with nearly correct predictions coming from another 16 groups. Several of the top scoring predictions come from fold recognition groups, owing to the similarities between this protein and DNA-binding three-helical bundles [e.g., 1au7; GeneSilico, 517; Fig. 7(D)], SAM-like domains (e.g., 1a0p; Celltech, 028), annexins (e.g., 1hm6; Pmodel3 server, 045), or the C-terminal domain of type II protein tyrosine phosphatases (e.g., 2shp, residues 461–525; Pushchino, 203). The Baker prediction is further noteworthy for a reasonable prediction of the N-terminal segment, which many of the others missed [Fig. 7(B)].

T0172 (NF/FR): MRAW from T. maritima *domain 2*

This protein consists of six helices, packing into an irregular structure. No group managed to predict the correct overall fold, although notable predictions included those from the Baker (002), Skolnick-Kolinski (010), and BioInfo.pl (006) groups who managed to predict parts of the structure correctly.

T0173 (NF/FR): Mycothiol deacetylase *from* M. tuberculosis

This structure consists of a doubly-wound, Rossmann-like, α/β structure with a parallel β-sheet of strand order 6-10-5-4-1-2-3 [Fig. 8(A)]. An insertion into this structure contains a separate meander of 3 β-strands, 7-8-9 [Fig. 8(C)].

Although no groups managed to reproduce the precise sheet topology, there were a number of good predictions. The similarity to other doubly-wound α/β folds meant it was unsurprising to see several successes come from

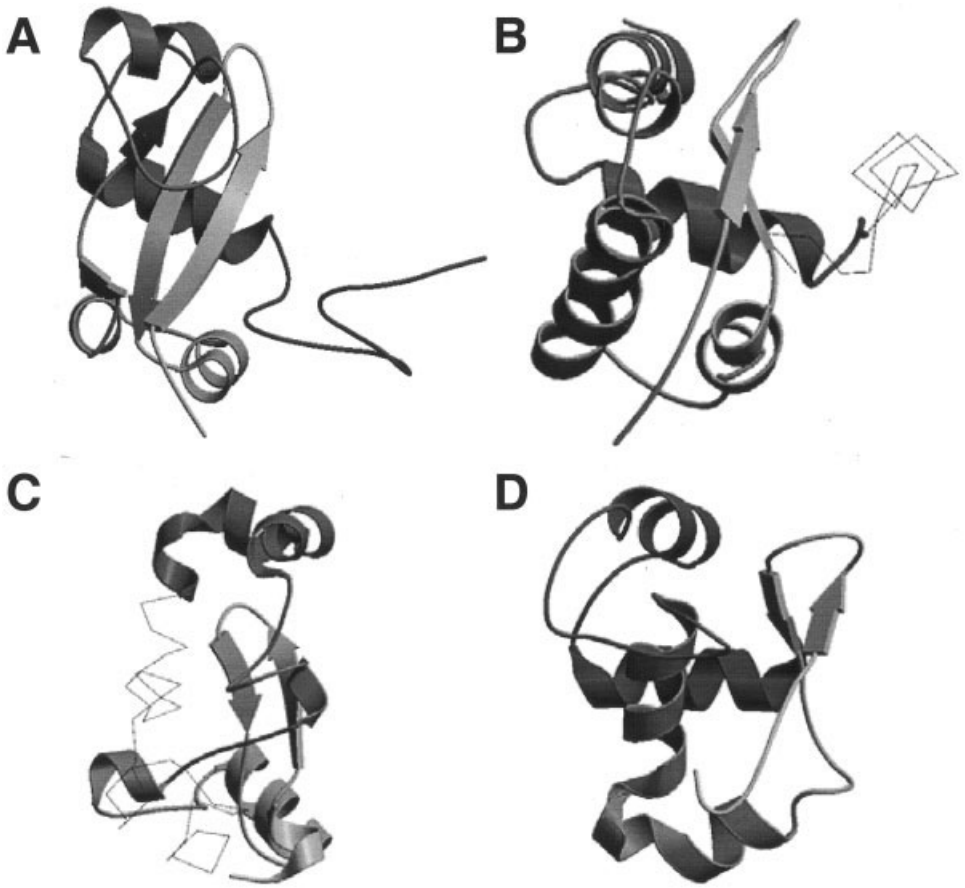

Fig. 5. Predictions for T0181. Correct structure (**A**) and predictions from the groups of (**B**) Skolnick-Kolinski (010, model 2), (**C**) SAM-T02-human (001, model 2), and (**D**) Samudrala-NewFold (051, model 3). Details are as for Figure 1.

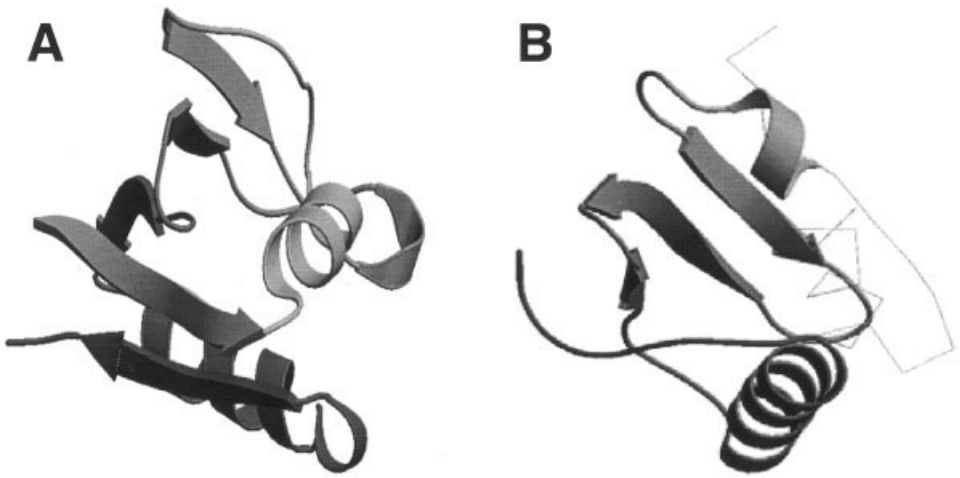

Fig. 6. A prediction for T0146_2. Correct structure (**A**) and prediction from the group of (**B**) Shortle (349, model 1). Details are as for Figure 1.

template-based predictions via fold recognition methods, including those from the Pmodel3 server (045) and from the Camacho [099; Fig. 8(B)] and Lomize (288) groups. However, other good predictions appear to have come from fragment-based de novo methods, such as those from the Baker (002) and Shortle (349) groups. The Baker group deserves special mention for predicting not only the N-terminal α/β structure well but also the C-terminal β-meander and helix [Fig. 8(D)], although the orientation of these subdomains differs from the X-ray structure [thus

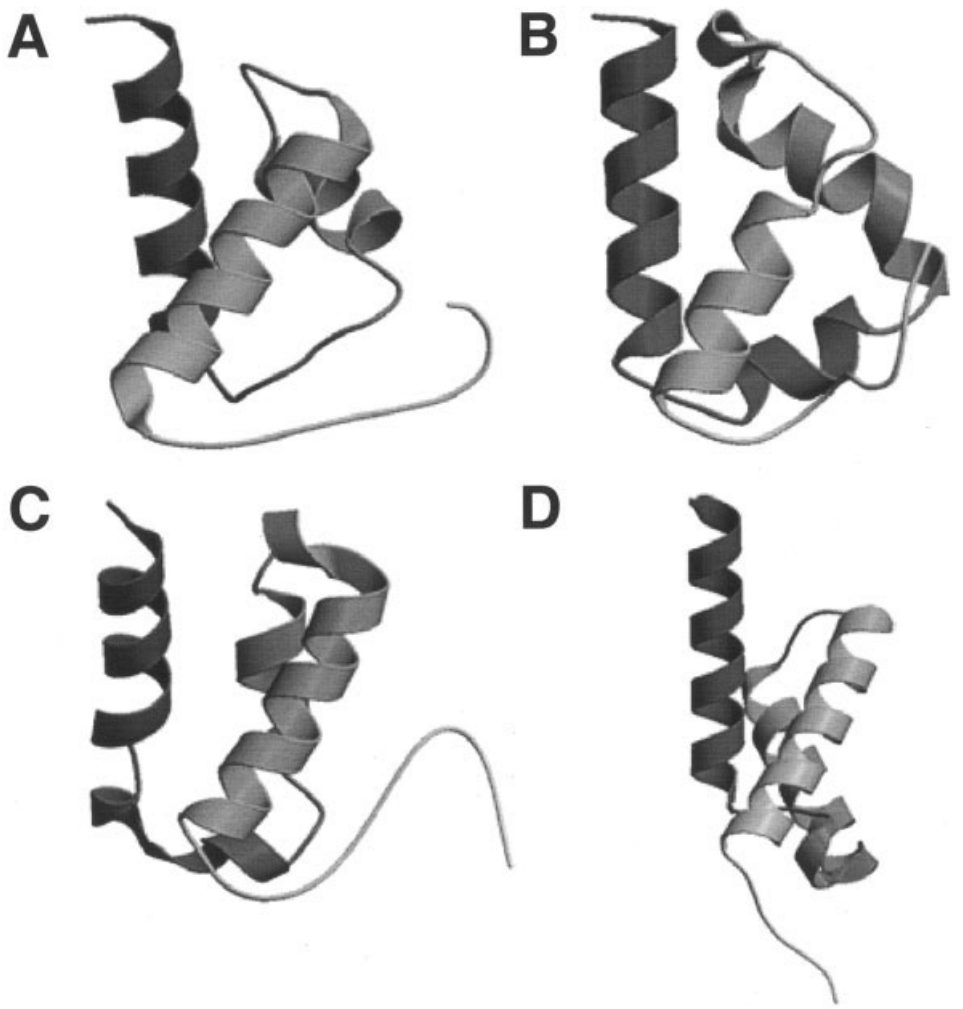

Fig. 7. Predictions for T0170. Correct structure (**A**) and predictions from the groups of (**B**) Baker (002, model 4), (**C**) Wolynes-Schulten (294, model 1), and (**D**) GeneSilico (517, model 1). That shown in (D) was based on a template (1au7). Details are as for Figure 1.

explaining the apparent differences in the Cα trace between Figs. 8(C) and (D)].

T0186 (NF/FR): TM0814 from T. maritima *domain 3*

This short domain inserted into an otherwise comparative modeling target was seemingly overlooked by many groups. We found only one correct prediction of fold, from the Baker group (002; Fig. 9). This prediction is spot-on, although as for T0173, its relative orientation to the other domains differs from that seen in the X-ray structure.

T0187 (NF/FR): TM1585 from T. maritima *domain 1*

This domain adopts an unusual, doubly-wound, α/β fold, comprising a six-stranded β-sheet with topology 126345, where strands 4 and 6 are antiparallel to the rest, and strand 5 is very short. No group predicted a correct topology, although parts of the structure were accurately predicted by the Libellula (230) and Baker (002) groups, the former using a purine repressor (2puf) as a template.

Overall view of performance

Tables II and III show views of the overall performance for groups making the best predictions by the visual and *z* score-based assessment, respectively. The precise ranking of groups depends on the weight given to the NF/FR targets where fold recognition groups were often able to get reasonable scores by identifying a new, but not exact match to a previously known fold. Although the visual assessment was quite ruthless, awarding points only to models close to the true structure, the tables show that the rankings broadly agree with those defined automatically by the *z* score. We found the other ranking schemes (see Materials and Methods) to be less effective, largely because they gave points for predictions that were clearly wrong (e.g., 50th percentile), meaning that groups with many average or wrong entries outperform those with just a few very good predictions. For most targets, the number of correct or good predictions is still comparatively small, arguing for

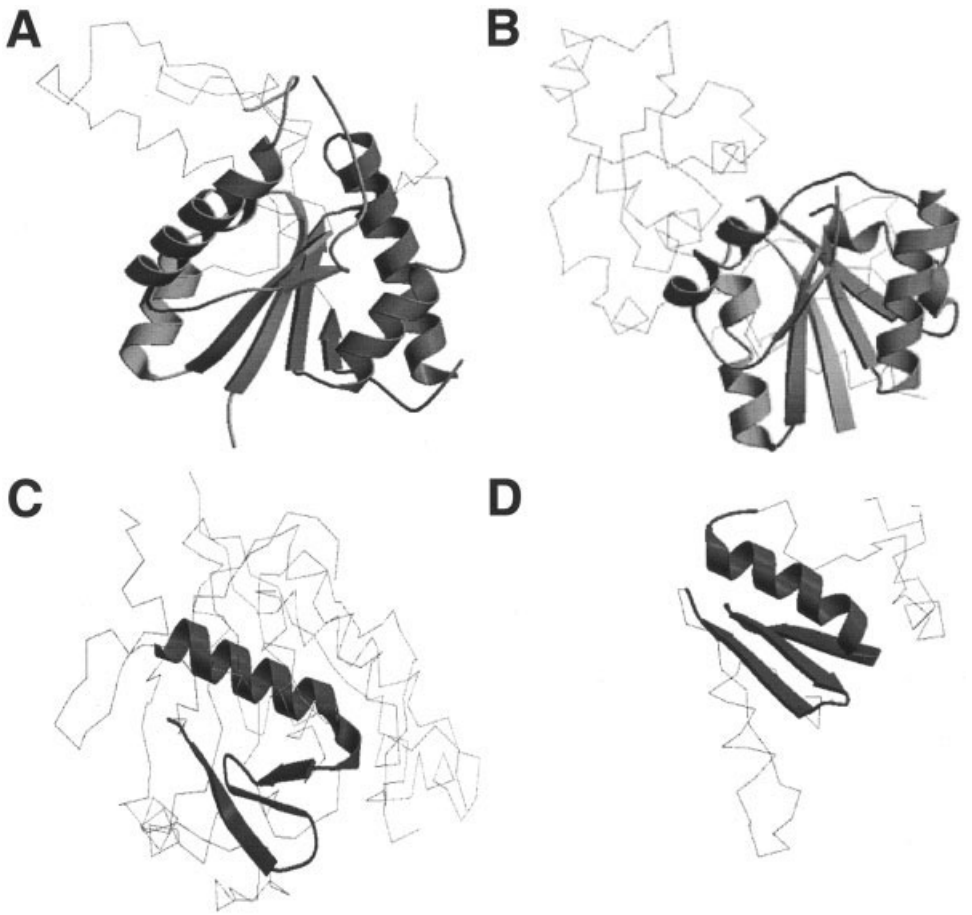

Fig. 8. Predictions for T0173. Correct structure (**A**) and predictions from the groups of (**B**) Camacho (099, model 2) and (**C**) Baker (002, model 1). (**B**) is based on a template (1f0k) and covers only the N-terminal, doubly-wound α/β portion of the molecule. The Baker group also predicted the N-terminal segment accurately (not shown), but they are further remarkable in that they have also accurately predicted the part of the C-terminus shown. Details are as for Figure 1.

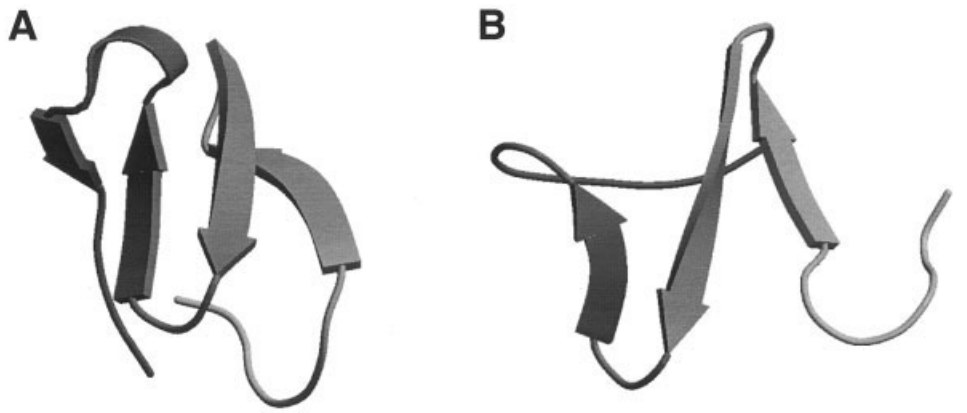

Fig. 9. Predictions for T0186_3. Correct structure (**A**) and prediction (**B**) from the Baker group (002, model 1). This is a small segment in an otherwise CM/FR target that no other group predicted with any degree of accuracy. Details are as for Figure 1.

a strategy like a z score, which awards most points for values high above the average.

Considering overall performance in both the visual and the z score-based assessment, the Baker (002), Jones-NewFold (068), and Shortle (349) groups stand out for the five new fold targets, with the SAM-T02-human (001) and Skolnick-Kolinski (010) groups also doing well when these and the NF/FR borderline targets are considered. The

TABLE II. Summary of Visual assessment

	α	α + β	α + β	α + β	α + β	NF		α + β	α + β	α + β	α	α	α/β	β	α/β	NF + NF/FR	
Group (No.)	129	149_2	161	162_3	181	T	S	146_1	146_2	146_3	170	172_2	173	186_3	187_1	T	S
Baker (2)	2 (4)	1 (1)	2 (2)	0	0	5	5	0	1 (5)	0	2 (4)	1 (5)	2 (1)	2 (1)T*	1 (2)	13	14
Shortle (349)	1 (1)	2 (2)	1 (1)	0	0	5	4	0	2 (1)	0	0	—	1 (1)	—	0	11	7
Jones-NewFold (68)	1 (4)	0	2 (3)	0	1 (4)	5	4	—	—	—	1 (2)	—	—	—	0	7	5
Skolnick-Kolinski (10)	0	0	0	0	2 (2)	5	2	0	0	0	2 (3)	1 (4)	0	0	0	13	5
S:I-sites/Bystroff (132)	0	0	0	2 (1)	0	5	2	1 (1)	0	0	0	0	0	0	0	13	3
Brooks (373)	0	0	0	2 (4)T	0	5	2	0	0	0	0	0	0	0	0	13	2
SAM-T02-human (1)	1 (3)	0	0	0	1 (2)	5	2	0	0	0	0	0	0	0	0	13	2
3D-PSSM (229)	0	0	0	1 (4)T	0	5	1	1 (5)T	1 (5)T	0	0	0	0	0	0	13	3
S:PROTINFO-AB (140)	0	0	0	0	1 (3)	5	1	0	0	—	2 (5)	0	0	—	0	11	3
Samudrala-NewFold (051)	0	0	0	0	1 (3)	5	1	0	0	0	2 (5)	0	0	0	0	13	3
ORNL-prospect (12)	0	0	0	0	1 (1)T	5	1	0	0	0	0	0	1 (1)T	0	0	13	2
Doniach (401)	1 (2)	—	—	—	0	2	1	—	—	—	0	—	—	—	—	3	1
Scheraga (314)	0	1 (2)	0	—	0	4	1	—	—	—	0	—	—	—	—	5	1
Head-Gordon (271)	—	0	0	1 (4)	—	3	1	—	—	—	0	—	—	—	—	4	1
Wolynes-Schulten (294)	—	—	0	—	0	2	0	—	—	—	2 (1)	—	0	—	—	4	2
Levitt (16)	0	0	0	—	0	4	0	0	1 (4)	0	1 (2)	0	0	0	0	12	2
GeneSilico (517)	0	0	0	0	0	5	0	—	—	—	2 (1)T	0	0	0	0	10	2
ATOME (464)	0	0	0	0	0	5	0	0	0	0	1 (5)T	0	0	1 (5)T	0	13	2
S:Pmodel3 (45)	0	0	0	0	0	5	0	0	0	0	1 (4)T	0	1 (2)T	0	0	13	2
S:Baker-Robetta (29)	0	0	0	0	0	5	0	0	1 (3)	0	1 (3)	0	0	0	0	13	2
BioInfo.PL (6)	0	0	0	0	0	5	0	0	0	0	0	1 (1)	0	0	0	13	1
Bilab (80)	0	0	0	0	0	5	0	0	0	0	1 (2)	0	0	0	0	13	1
Rokko (327)	0	0	0	—	0	4	0	0	0	0	1 (3)	0	0	—	0	11	1

T denotes predictions made by using templates.
Shaded boxes show predictions awarded 2 points.
*For this prediction, the template specified was used for the other domains (D. Baker, personal communication). Groups who scored only 1 point using a template (T) were removed for clarity. A complete table is available at the accompanying Web site.

TABLE III. SUMMARY OF GDT_TS *Z* Score-Based Assessment

	α	α + β	α + β	α + β	α + β	NF		α + β	α + β	α + β	α	α	α/β	β	α/β	NF + NF/FR	
Group	129	149_2	161	162_3	181	T	S	146_1	146_2	146_3	170	172_2	173	186_3	187_1	T	S
Baker (002)	5 (4)	2 (1)	3 (2)	2 (2)	2 (2)	5	14	1 (2)	2 (5)	1 (2)	4 (4)	3 (5)	4 (1)	6 (1)T*	2 (2)	13	37
Jones-NewFold (068)	3 (4)	2 (5)	3 (3)	2 (2)	2 (4)	5	12	—	—	—	2 (2)	—	—	—	1 (5)	7	15
Skolnick-Kolinski (010)	2 (1)	2 (4)T	1 (2)T	1 (2)T	5 (2)	5	11	1 (1)	1 (1)	1 (1)	4 (3)	2 (2)T	1 (4)	1 (1)T	2 (4)T	13	24
SAM-T02-human (001)	2 (3)	1 (2)	2 (1)	2 (1)	4 (2)	5	11	1 (1)	2 (5)	1 (3)	1 (3)	1 (4)	1 (1)	0 (1)	2 (5)	13	21
Shortle (349)	2 (1)	2 (2)	2 (1)	1 (1)	1 (2)	5	9	1 (1)	4 (1)	1 (1)	0 (1)	—	3 (1)	—	2 (2)	11	19
Brooks (373)	1 (3)	1 (2)T	2 (2)	3 (4)T	2 (2)	5	8	1 (2)T	1 (1)T	1 (1)T	1 (1)	0 (3)T	1 (4)T	1 (1)T	1 (3)T	13	15
S:Baker-Robetta (029)	2 (3)	1 (1)T	2 (3)	2 (1)	1 (1)T	5	8	1 (5)	3 (3)	1 (5)	2 (3)	1 (5)T	0 (5)	0 (1)T	2 (2)	13	18
Scheraga-Harold (314)	2 (1)	2 (2)	2 (4)	—	1 (5)	4	7	—	—	—	1 (5)	—	—	—	—	5	7
Samudrala-NewFold (051)	1 (5)	2 (2)	1 (3)	1 (3)	3 (3)	5	7	1 (5)	2 (4)	1 (4)	3 (5)	1 (1)	0 (1)	0 (5)	1 (3)	13	16
S:PROTINFO-AB (140)	1 (5)	2 (2)	1 (3)	1 (3)	3 (3)	5	7	1 (5)	1 (5)	—	3 (5)	1 (1)	0 (4)	—	1 (3)	11	15
Levitt (016)	1 (1)	2 (3)	2 (5)	—	1 (1)	4	7	2 (4)	2 (4)	2 (1)	2 (2)	2 (5)T	1 (1)	0 (1)T	2 (2)	12	21
S:I-sites/Bystroff (132)	0 (5)	2 (1)	2 (1)	3 (1)	1 (1)	5	7	2 (1)	1 (1)	0 (1)	0 (1)	2 (1)	0 (1)	1 (1)	1 (1)	13	14
Head-Gordon (271)	1 (4)	1 (5)	0 (2)	2 (2)	1 (4)	5	6	1 (4)	2 (4)	1 (3)	1 (1)	—	—	—	1 (2)	10	11
KIAS (531)	1 (1)	—	3 (1)	2 (1)	0 (1)	4	5	—	—	—	1 (5)	2 (1)	—	—	—	6	8
keasar (429)	—	2 (3)	—	—	2 (5)	2	4	—	—	—	1 (1)	1 (1)	1 (2)	—	—	5	7
Rokko (327)	1 (1)T	1 (1)T	0 (2)T	—	2 (3)	4	4	0 (1)T	0 (1)T	0 (1)T	2 (3)	0 (2)T	0 (2)	—	1 (1)T	11	7
Ginalski (453)	0 (1)	1 (1)T	1 (1)	0 (1)	1 (1)	5	3	1 (1)	0 (1)	0 (1)	1 (1)	2 (1)	3 (1)T	1 (1)T	1 (1)	13	12
Doniach (401)	2 (2)	—	—	—	1 (3)	2	3	—	—	—	1 (3)	—	—	—	—	3	4
BioInfo.PL (006)	0 (1)	1 (1)	1 (1)	0 (1)	1 (1)	5	3	1 (1)	0 (1)	0 (1)	1 (1)	2 (1)	1 (1)	1 (1)	1 (1)	13	10
Wolynes-Schulten (294)	—	—	1 (2)	—	2 (1)	2	2	—	—	—	2 (1)	—	0 (1)	—	—	4	5
Samudrala-FR (052)	0 (5)	0 (2)	0 (1)	1 (1)	0 (2)	5	2	1 (4)	0 (1)	1 (1)	0 (1)	1 (5)	0 (5)	2 (5)	0 (3)	13	7
S:PROTINFO-FR (139)	0 (3)	0 (2)	0 (1)	1 (1)	0 (2)	5	2	1 (4)	0 (1)	1 (1)	0 (1)	1 (5)	0 (5)	2 (5)	0 (3)	13	6
Avbelj-Franc (341)	—	—	—	—	2 (3)	1	2	—	—	—	0 (4)	—	—	—	—	2	2

To be included in this table, groups had to have attained a *z* score of 2 or higher for at least one prediction without using a template. A complete table is available at the accompanying Web site. Other details are as for Table II.
Shaded boxes show predictions with *z* score ≥ 3.

Baker group stands well apart from the others when all targets are considered, and indeed, the reader will notice that they often feature among the predictions that accurately identify particularly features.

We inspected differences between the visual and *z* score assessment (Table II compared to Table III) and found that they were either due to methods where only fragment prediction, and not assembly, was accurate, or where predictions contained overlapping segments (see Limitations of numerical evaluations section below). Correct fragments often score similarly by GDT_TS to correct overall folds, and our preference for overall fold meant that groups only predicting fragments were down-weighted in the visual relative to the numerical evaluations. One such group had spent most of its effort developing a sophisticated fragment library and was still using an experimental fragment assembly algorithm during CASP5 (Kevin Karplus, personal communication).

On the whole, predictions from servers performed less well than those where human intervention was used. The small numbers and similarity in performance make it difficult to determine which of PROTINFO-AB (149), I-sites/Bystroff (129), or BAKER-ROBETTA (029) performs best. It was also clear that the visual assessment tended to down-weight them. Inspection showed that, like the predictions above, the servers were better at predicting local fragments than assembling them. This suggests that fragment assembly is where human input reaps the most benefit.

Selection of best model

Predictors are encouraged to select one model to be a "best" model at rank 1. Certain groups, including those of Baker (002) and Shortle (349), continue to perform well in both the visual and automated assessments, although these and others perform less well overall than when considering the best model submitted. Equivalents to Tables II and III where only 1st models are considered are given in the accompanying Web site.

How were good predictions made?

For a selection of the top scoring groups from Tables II and III, we solicited feedback via e-mail to produce the methods summary shown in Table IV (the full list of questions and a more detailed table appear in the accompanying Web site). A few trends stand out. Nearly all these methods make use of secondary structure prediction (20 of the 25 groups use PSIPRED[16]). More than half use homologous sequence information, typically for secondary structure prediction or as input for threading. The groups are split roughly equally between automated methods and methods allowing some kind of manual intervention. Manual intervention is most often used to choose templates or fragments and to inspect models. Only nine groups considered their method to be "exclusively fold recognition." The remaining are fragment based or ab initio methods, or methods that combine several approaches depending on the type and difficulty of the target. At least seven groups appear not to have used either fragments or templates to generate their predictions.

Some predictions stand out as successes for "analogous" fold recognition, such as those by the Brooks group (373) for T0162_3, the GeneSilico group (517) for T0170, and several groups for the Rossmann-like structure in T0173. However, fragment-based methods perform better overall on true FR analogs,[17] arguing that they are perhaps better used to predict targets for which remote homology cannot readily be inferred. It is interesting to speculate how the performance of groups who chose to use only FR methods on certain targets might have altered if de novo methods had been applied instead.

Comparison with previous CASPs

It is difficult, if not impossible, to compare the results here to those from previous CASP experiments, owing to differences in methods, assessment philosophies, and target classification or difficulty. Nevertheless, we tried a number of schemes to see if we could detect progress in the ability of the community to predict structures for proteins adopting new folds.

First, we calculated the average GDT_TS for the top three predictions per target for CASP4 and CASP5. We found a slight difference between the two experiments: an average of 31.4 in CASP4 and 33.8 in CASP5, suggesting a moderate improvement. However, high standard deviations (13.0 and 11.9) make the difference statistically insignificant. The GDT_TS values largely depend on the targets; thus, the results are not easily comparable.

We also tried to compare our visual assessments. To do this, we inspected the predictions from CASP4 in a similar manner and checked whether we would have awarded 2, 1, or 0 points to *any* prediction. For 13 CASP5 targets, there were 9 with at least one prediction getting 2 points, with an additional 3 targets getting 1 point (no points were given for T0146_3). For 14 (of 17) CASP4 targets, where assessment data were available, the equivalent numbers are 8 and 6. This simple comparison suggests only minor overall progress (i.e., 9 of 13 compared to 8 of 14). However, a greater impression of progress comes when one ignores the NF/FR (borderline) targets. For all five of the CASP5 true NF targets, at least one prediction was awarded 2 points, which is true for only one of the five of those from CASP4.

Limitations of numerical evaluations

Differences between the visual and numerical rankings revealed instances where comparatively high GDT_TS scores were achieved by what we considered to be wrong predictions. We present these here largely for the benefit of future assessments.

Long helices can lead to high GDT_TS scores despite differences in overall structure. For example, a 37-residue C-terminal helix in the $\alpha + \beta$ T0162_3 resulted in instances of higher GDT_TS values for predictions of very different structures (e.g., all-α-proteins) that contained a single long helix, or coiled-coil segment [e.g., Fig. 10(A)] than for those that had identified arguably more useful features of the structure such as all five strands of the core β-sheet. α/β proteins also present problems: we noticed

TABLE IV. Short Summary of Methods Used in CASP 5

Group Name	Group no.	Only FR?	Altered method?	Sec Str?	Manual steps?	Homologues?	Domains?	Fragment-based?	Other servers?	Lattice-based?	Threading-potentials?	Optim?
SAM-T02-human	001	N	N	Y	Y	Y	N*	Y	Y	N	Y	N
BAKER	002	N	Y	Y	Y	Y	Y	Y	Y	N	Y	Y
Skolnick-Kolinski	010	N*	N	Y	N	Y	N*	N	N	Y	Y	Y
ORNL-PROSPECT	012	Y	N	Y	Y	Y	Y	N	Y	N	Y	N
S:BAKER-ROBETTA	029	N	Y	Y	N	Y	Y	Y	Y	N	Y	N
S:Pmodel	040	Y	N	Y	N	N	N	N	Y	N	N	N
S:Pmodel3	045	Y	N	Y	N	N	N	N	Y	N	N	N
Samudrala-NewFold	051	N	N	Y	N	N	Y	Y*	N	N	Y	Y
Jones-NewFold	068	N	N	Y	N	Y	N	Y	N	N	Y	N
Friesner	112	N	Y	Y	Y	Y	Y*	N	N	N	Y	Y
S:I-Sites/Bystroff	132	N	Y	Y	N	Y	N*	Y	N	N	N	Y
S:PROTINFO-AB	140	N	N	Y	N	N	Y	Y*	N	N	Y	Y
FAMS	168	Y	N	Y	N	N	N	N	N	N	N	Y
ORNL-PROSPECT	195	Y	N	Y	N	Y	N	N	N	N	Y	N
Pushchino	203	N	N	Y	Y	Y	Y	N	Y	N	Y	N
3D-PSSM	229	Y	N	Y	N	Y	N	N	N	N	Y	N
Head-Gordon	271	N	N	Y	Y	N	N	N	Y	N	N	Y
Wolynes-Schulten	294	N	N	Y	Y	Y	Y	N	Y	N	Y	Y
Scheraga-Harold	314	N	N	Y	Y	N	Y	N	Y	N	N	Y
Shortle	349	N	N	Y	Y	Y	N*	Y	Y	N	Y	N
Brooks	373	N	Y	Y	Y	N	Y*	N	Y	Y	N	Y
Doniach	401	N	N	Y	Y	N	Y	N	Y	N	N	N
TOME	450	Y	N	Y	Y	Y	Y*	N	Y	N	Y	N
ATOME	464	Y	N	Y	N	Y	N	N	Y	N	Y	N
GeneSilico	517	Y	N	Y	Y	Y	Y	Y	N	N	N	N

Selected groups were asked to provide more detailed information about their methods. The full list of questions can be found in the accompanying Web site. Y/N values marked with an asterisk (*) indicate answers that are somewhat ambiguous and better described in a more extensive table also found on the same Web site. In short, we asked predictors the following: whether they considered their method exclusively fold recognition; whether they altered the method for different NF or NF/FR targets; if the method incorporated predicted secondary structure information, included any manual intervention, used homologous sequence information, or defined domains before prediction; whether the method was fragment-based; whether any other publicly available servers were used as part of the prediction process; if the method included lattice-based representations of protein structures, incorporated threading-like potentials, or if it included any steps for relaxation, minimization, or optimization of the models.

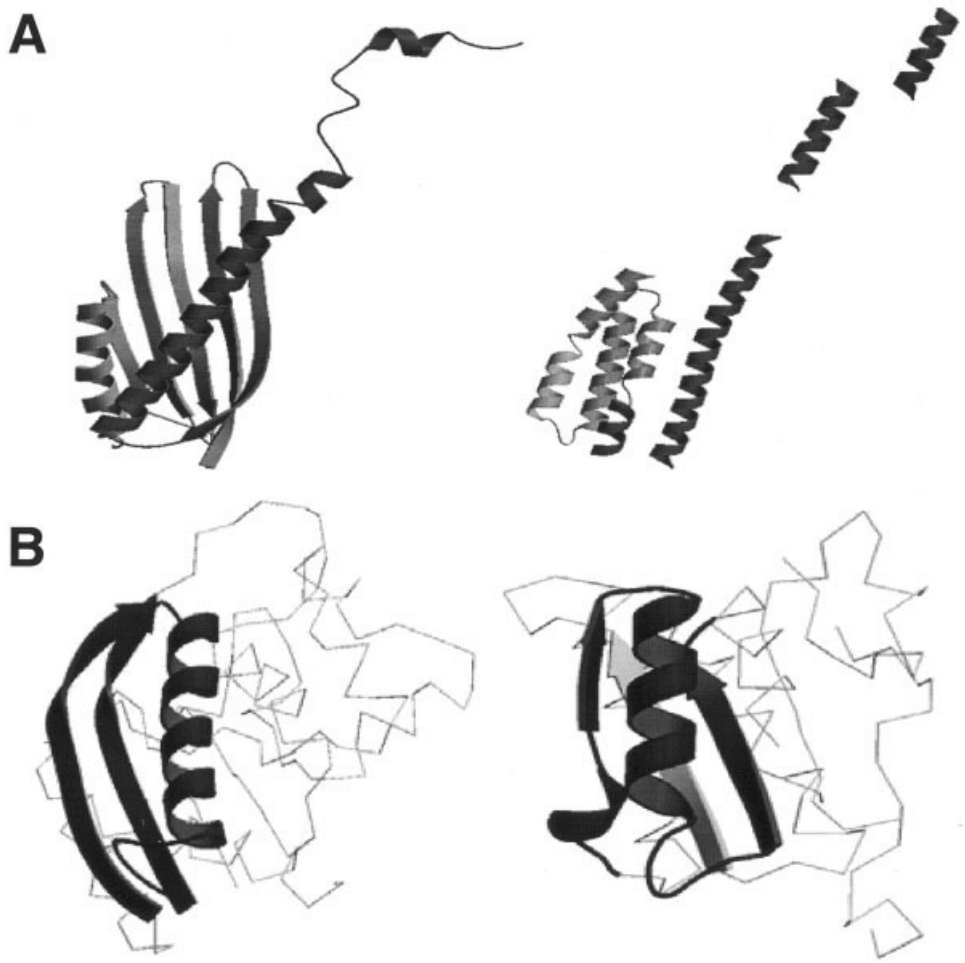

Fig. 10. Examples of limitations of GDT_TS. **A:** An all-α-prediction (right) scores well with an α+β target (T0162_3, left) owing to a single long α-helix. **B:** An example of topologically different β-α-β segments (dark color) that give points in GDT_TS. The predicted β-α-β unit (right) has a β-strand inserted into it (light color).

several instances where β-α-β units from sheets with different numbers of strands lying between them led to high GDT_TS [e.g., Fig. 10(B)].

We also observed that several predictions, which appeared to comprise overlapping segments, ranked comparatively highly when assessed by using GDT_TS, but did poorly in the visual assessment. After uncovering limitations in the program that calculated GDT_TS (where singularities gave artificially high scores; Adam Zemla, personal communication), we discovered that it was often possible to improve GDT_TS values (with increases of ≥8) by simply cutting predictions at arbitrary positions, overlaying the sequences, and using whatever equivalences resulted to superimpose the fragments. It appears that when segments are superimposed, they can profit from GDT 8 and 4 Å distance components,[12] which would be less likely if the segments were assembled into a plausible, yet incorrect, structure.

Avoiding situations where anomalies score highly is particularly important if performance is ultimately to be assessed automatically. Thus, we would suggest exploring variations in GDT_TS in future CASPs. A means to give different weights to segments according to secondary structure type might avoid the problems above, and it might be prudent to consider modifying the calculation to avoid wrong topologies getting artificially high scores, particularly for β-sheets [e.g., Fig. 10(B)]. We also suggest further investigation into the phenomenon of overlapping coordinates to avoid problems in future CASPs. We attempted to account for this by introducing a filter to remove predictions with many close Cα–Cα contacts. Table III contains the unfiltered ranking, but filtered rankings are available on the accompanying Web site.

Secondary Structure Predictions

For this assessment we considered all 54 targets, split into 78 domains.[15] Division of targets into domains is logical because some targets contain domains from different categories (e.g., T0149 CM and NF) and, therefore, different prediction difficulties. A total of 38 groups, including 10 servers, submitted a total of 2626 predictions in this category. Like CASP4, we considered only model 1 during the assessment.

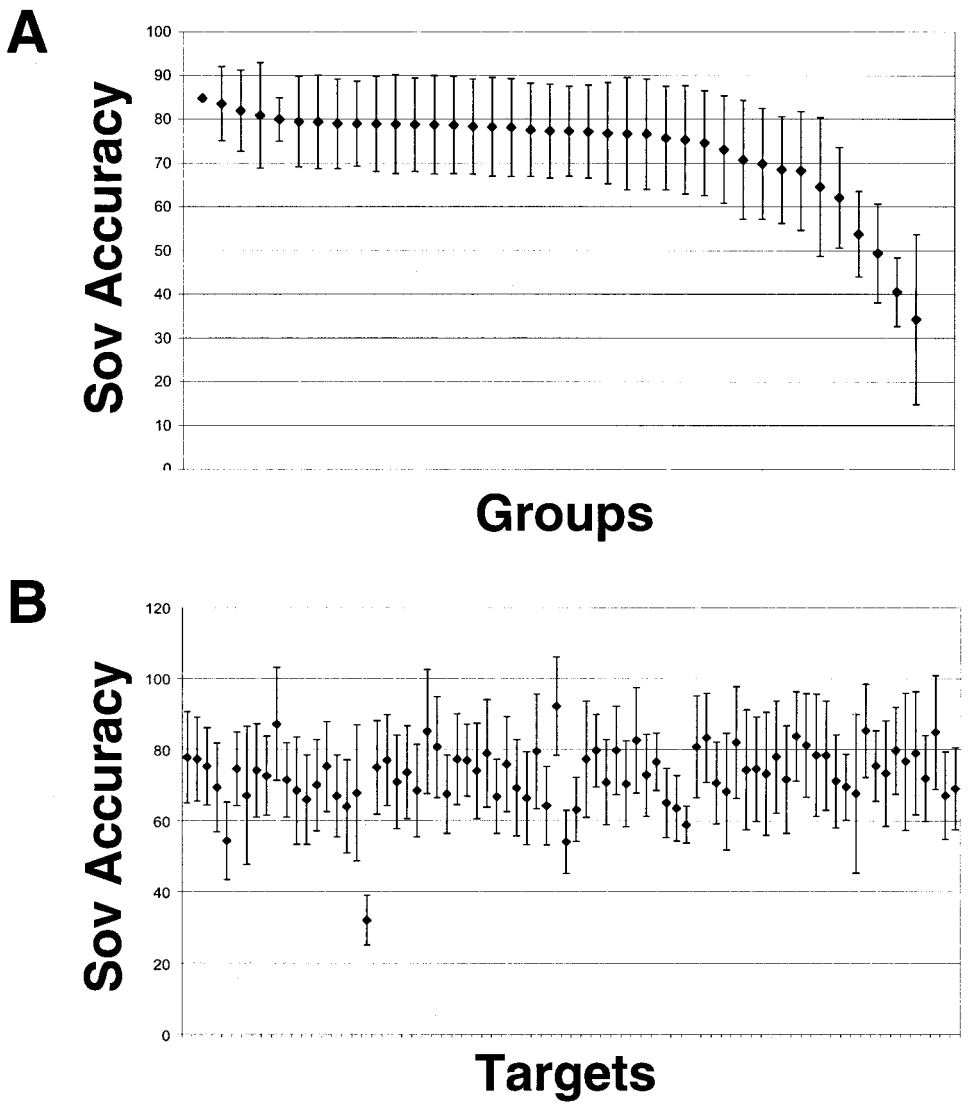

Fig. 11. SOV means and standard deviations for each group (**A**) and target (**B**).

We assessed predictions using segment overlap (SOV[11,18]) and three-state per residue (Q3) measures. Here we consider only SOV because it is widely accepted to be a better means to assess secondary structure prediction accuracy.[4,18] Its superiority is apparent when considering predictions for the all-α-protein T0129, where several near-perfect predictions are correctly top ranked according to SOV, but the best according to Q3 predicts a β-strand between residues 80 and 85. We observed no major differences in the results when we substituted SOV with Q3. All results are available in the accompanying Web site.

As argued during the CASP 4 assessment,[4] SOV or Q3 mean and standard deviations are not effective as a means to rank group performances over a variety of targets. Many groups had very similar mean values with large standard deviations, which makes it impossible to distinguish between them [Fig. 11(A)]. Moreover, with the normalization implicit in any calculation of a mean, a group submitting only one accurate prediction (e.g., ORNL-PROSPECT, 195), and not necessarily the best one for that particular target, would get a higher score than those groups submitting consistently good predictions for all the targets (e.g., PSIPRED, 072). Thus, we needed a scoring system able to rank groups based on relative performance. We decided to use the scheme used during CASP4[4] based on SOV percentiles (see Materials and Methods).

Figure 11(A) shows that the same group can achieve very different accuracies for different targets. Moreover, averages and standard deviations of the prediction performance for each target [Fig. 11(B)] reveal that some targets vary greatly in prediction difficulty. This finding suggests a highly variable intrinsic prediction difficulty, which prompted us to classify the targets accordingly (see Materials and Methods): 28 were classified as easy, 46 as medium, and 4 as hard. In CASP4, the equivalent numbers were 14 easy, 33 medium, and 11 hard. Ignoring the possibility that target difficulty might be different in the two experiments, the classification difference suggests

that the field has improved: easy targets have increased and hard have decreased. Unfortunately, we could not use this classification further because too few targets were considered hard and any derived result would be statistically insignificant. Moreover, the prediction accuracies for one of the hard targets (T0146_3) were probably artificially low because of disruptions in the hydrogen-bonding pattern in the crystal structure that led DSSP[19] to identify only two of four likely β-strands in the core β-sheet. Nevertheless, the calculations based on this classification are available as additional information in the accompanying Web site.

In the end, we opted for a simple division of targets into sequence related (SR) and sequence unrelated (SU). The first group contains the 57 target domains that show sequence similarity to a protein of known three-dimensional (3D) structure or convincing evidence for a common ancestor [CM, CM/FR(H), and FR(H)]. The second contains 21 target domains without detectable sequence similarity or additional evidence supporting a common ancestor [FR(A), FR(A)/NF and NF]. Note that the different numbers of targets in the two categories will clearly bias the results for all predictions toward sequence-related proteins.

Figure 12 shows plots of several overall performances achieved in CASP5. We considered only the top 10 scoring groups for each subset of targets. Figure 12(A) shows the overall accuracies seen in CASP5 compared to CASP4. For all targets, the overall accuracy increased from 76.1 to 79.9%, with an average increase in the performance of the best predictor by nearly 5%. However, when considering the SR and SU subsets separately, the improvement is much less. This might be due to the fact that in CASP4, there were roughly equal numbers of SR (30) and SU (28) targets, whereas for CASP5, the number of SR targets (57) is almost three times that for SU (21).

β-strands have historically been more difficult to predict than α-helices, and we observe that this is still the case for targets in CASP5 [Fig. 12(B)]: helices are predicted on average 9.5% more accurately than strands. This effect is more pronounced when we compare only the SU targets (20.5%).

We also compared the performance of the 10 fully automated servers to the human predictors [Fig. 12(C)] and found that servers are 6% better on average than humans when considering all predictions. However, the 10 best human predictors score 3% higher than servers. There is no difference between humans and servers in the SU subset, suggesting that this improvement is likely coming from good comparative models that were used to assign secondary structures.

Table V shows scores and SOV means for the top 10 groups in the three target subsets. Considering all or just SR targets, the Bujnicki-Janusz (020), CaspIta (108), and GeneSilico (517) groups stand out from the others. Bujnicki-Janusz (020) and GeneSilico (517) are related predictors that used similar strategies: they indeed built homology models (see Ref. 20) and used them to assign secondary structures. It is worth noting that this procedure was the only one able to predict all the secondary structure elements for target T0129 (SOV = 96.7), surprisingly, despite an incorrect prediction of fold. CaspIta (108) is a very successful metaserver, which combines the outputs of PSIPRED,[16] ssPRO,[21] and Sam-T2K.[22] It is interesting that when it detects sufficient sequence similarity to a known structure, it also derives secondary structures from a comparative model. The consensus obtained and some clever decisions in case of ties made them also the best for SU targets.

The situation is slightly different when we consider only SU targets. Here the groups deriving secondary structures from 3D models (i.e., Bujnicki-Janusz, 020) obviously disappear and some new groups [e.g., Kim-Park (442)], appear in the top 10. It is clear from these results that all the rankings are dominated by three original methods: PSIPRED (072), SAM-T02-server (189), and Baldi-SSpro (023), or modifications/combinations of their outputs such as CaspIta (108), MacCallum (393), and APSSP2 (055).

There is a 4% difference in SOV between the SR and SU categories, and the standard deviations are larger for SU targets. The small number of SU targets means that the apparently missed strands in T0146_3 (see above) accounts for 2.5% of this difference. We suspect that the remaining 1.5% is due to higher accuracies obtained by groups using comparative models.

Residue–Residue Contact Predictions

Here we considered the same 78 domains as for secondary structure predictions. Only six groups entered this category, submitting a total of 275 predictions. We also split the targets into sequence related (SR) and sequence unrelated (SU). Like previous CASPs, we considered only the first model submitted.

Because of the limited number of groups, we did not consider it necessary to develop a detailed process for ranking group performance; instead, we simply calculated predictive accuracy and coverage as described in Materials and Methods. In brief, the predictive accuracy shows how many of the predicted contacts really are in contact in the target structure. The predictive coverage reports how many of the observed contacts have been correctly predicted. Individual per target values of accuracy and coverage are available in the accompanying Web site.

To make assessment easier, groups were encouraged to submit predictions of contacts between residues within 8 Å; however, only the Baldi-CONpro group (022) chose to use a different cutoff (12 Å). For simplicity, we defined two residues (R1 and R2) to be in contact if the distance between their Cβ atoms (Cα for glycine) was ≤8 Å. Following CASP guidelines (http://PredictionCenter.llnl-.gov/casp5), we also divided contacts into three different categories depending on sequence separation: short (R1,R2 are ≤4 residues apart), medium (5 ≤ R1,R2 ≥ 8), and long range (R1,R2 ≥9).

Table VI shows the overall accuracy and coverage results achieved by the six different groups. Considering overall accuracies, the Bujnicki-Janusz (020) and GeneSilico (517) groups performed much better than the

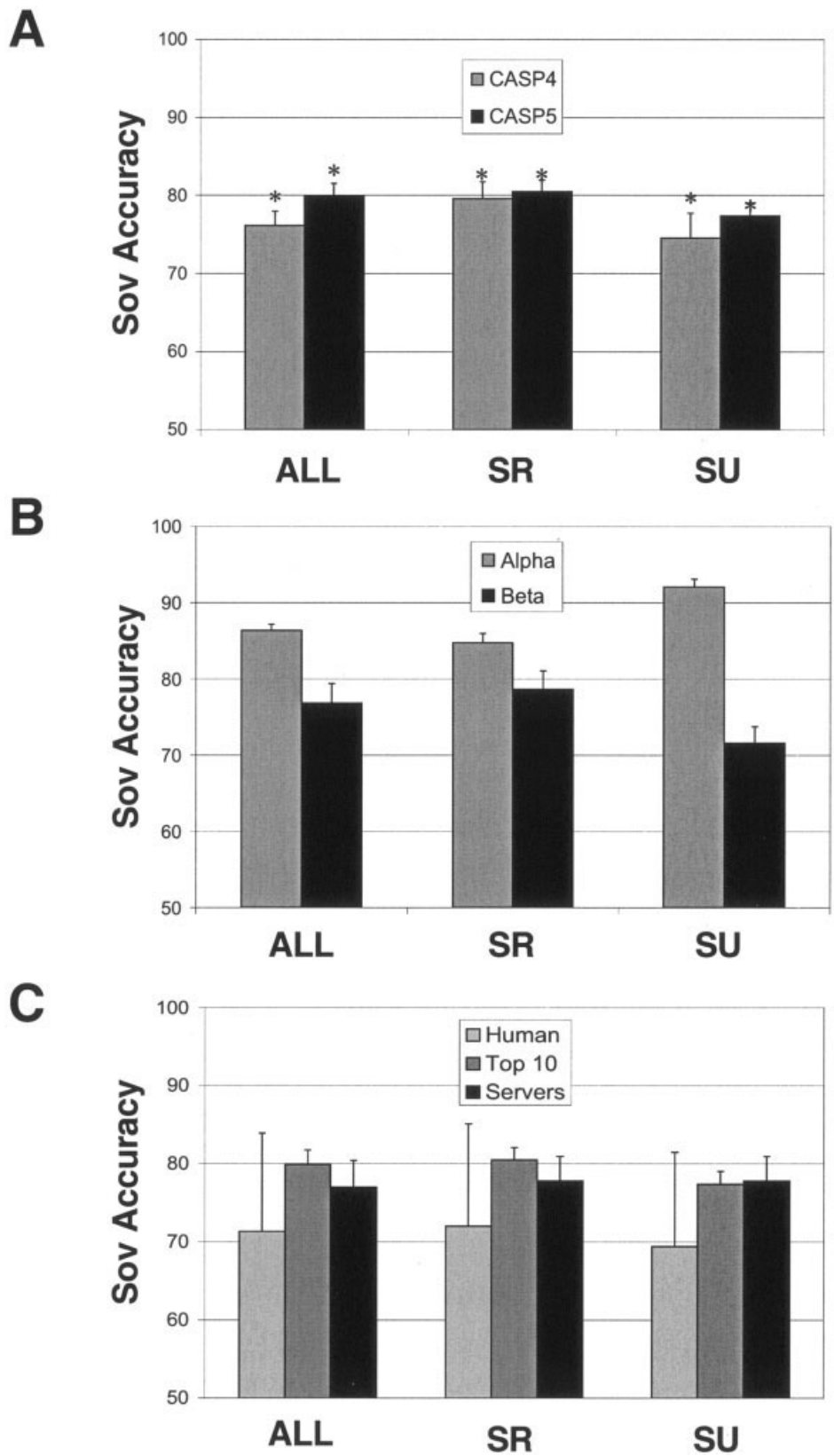

Fig. 12. Secondary structure prediction performances. Comparison of the overall accuracies for the top 10 scoring groups for the following: CASP5 vs CASP4 (**A**); Alpha vs Beta (**B**), and Servers vs Humans (**C**).

TABLE V. Secondary Structure Prediction Results for the Top 10 Scoring Groups

All			Sequence Related			Sequence Unrelated		
Group	Score	⟨Sov⟩	Group	Score	⟨Sov⟩	Group	Score	⟨Sov⟩
020	51.5	83.5 ± 8	020	58.8	83.3 ± 9	108	50.8	78.3 ± 16
108	50.2	80.8 ± 12	517	55.0	82.2 ± 9	023	50.8	77.8 ± 15
517	49.8	81.9 ± 9	108	50.0	81.7 ± 10	442	44.4	77.2 ± 16
072	40.0	79.4 ± 11	072	41.5	80.3 ± 8	055	43.6	76.8 ± 14
068	39.3	79.4 ± 10	068	39.2	80.0 ± 8	001	42.8	76.2 ± 15
001	39.3	78.7 ± 11	393	38.6	79.8 ± 9	068	39.7	77.7 ± 15
023	38.5	78.9 ± 11	001	38.0	79.7 ± 8	254	38.9	74.9 ± 15
393	37.8	78.9 ± 11	389	35.7	79.3 ± 8	517	35.7	81.1 ± 11
055	37.2	78.3 ± 11	055	34.8	78.8 ± 9	389	35.7	77.7 ± 14
389	35.7	78.9 ± 10	023	33.9	79.3 ± 9	393	35.7	76.4 ± 16

For each independent set of targets [All, sequence related (SR), and sequence unrelated (SU)], the first column shows the group number, the second the total score for each group, and the third column the SOV mean and standard deviation.

TABLE VI. Contact Prediction Results

	Short range			Middle range			Long range		
Group	Np	Acc	Cov	Np	Acc	Cov	Np	Acc	Cov
020	27261	0.78	0.51	2991	0.42	0.06	22415	0.46	0.10
022	49993	0.60	0.72	34606	0.08	0.14	129663	0.07	0.09
131	3949	0.61	0.06	2668	0.17	0.02	10480	0.16	0.02
397	33596	0.78	0.63	6562	0.03	0.01	5654	0.16	0.01
517	27668	0.76	0.50	3112	0.36	0.05	19396	0.44	0.09
587	0	—	—	3838	0.06	0.01	129450	0.02	0.03

The first column shows the group number. For each particular subset of contacts (short, medium, and long range), the first column shows the total number of predicted contacts and the second and third columns show the prediction accuracy and coverage, respectively.

others. However, as for secondary structure prediction (see above), their strategy was to build comparative models and use them to derive the contacts. We did not think that this approach was comparable to the others, where to our knowledge homologous template information was not used, and contacts were predicted de novo.

For short-range contacts, several groups attained good accuracy and coverage. However, we do not think that this is a very difficult problem because most of the ±4 adjacent residues in sequence are within the 8 Å cutoff. We would also question the use of such contact predictions, other than perhaps another means to identify secondary structures.

As might be expected, both accuracy and coverage drop dramatically when we consider the medium-and long-range contacts, and the levels achieved are far from ideal. Curiously, the CIRB (397) group did significantly better at predicting contacts between residues far apart in the sequence than for those found in the medium range category. Baldi-CONpro (022) submitted about an order of magnitude more predictions than the other groups predicting contacts de novo; however, the accuracy achieved was average at best.

Groups were encouraged to submit reliability scores associated to each predicted contact; five groups followed this suggestion. The thousands of residue–residue contacts within a protein domain make an accurate prediction of all of them virtually impossible using de novo methods. Ideally, what one would like are a few reliable contacts rather than many inaccurately predicted.

If one considers only those predictions marked as reliable at ≥0.8, the accuracy improves considerably. Even ignoring groups using comparative models (e.g., Bujnicki-Janusz, 020; GeneSilico, 517), there are still some very accurate predictions (see accompanying Web site). For instance, the Baldi-CONpro (022) group was able to predict a few contacts for targets T0129, T0181, and T0186_3 with 100% accuracy. The CIRB (397) group also predicted some targets with remarkable accuracy. This success prompted us to investigate further the type of structural features identified by the predictors.

Table VII shows details of the accuracies and sequence separation for the long-range contact predictions. Considering all predictions, the average and standard deviation of the residue–residue sequence separation roughly covers the whole domain length and is comparable for all groups. Differences arise when we consider only reliable contacts predicted by de novo methods (e.g., Baldi-CONpro, 022; CIRB, 397). Although prediction accuracy increases for both groups, the Baldi-CONpro group (022) only predicts contacts ≤ 13 residues apart in sequence; (i.e., just inside the limit for the definition of long-range interactions). Inspection showed that most correct predictions were those between parts of the same supersecondary structures. In contrast, predictions from the CIRB group (397) still span the whole length of the protein and are thus more

TABLE VII. Sequence Separation for Long-Range Predicted Contacts

	Long range-all				Long range-reliability ≥ 0.8			
Group	Ncp	Acc	R_1-R_2⟨d⟩	R_1-R_2 SD	Ncp	Acc	R_1-R_2 ⟨d⟩	R_1-R_2 SD
020	366	0.52	34.3	25.0	59	0.64	31.5	6.9
022	1818	0.08	42.4	51.5	43	0.33	11.0	2.2
131	380	0.16	28.7	22.9	380	0.16	28.7	22.9
397	146	0.14	58.1	44.1	23	0.12	74.2	32.2
517	334	0.44	29.5	18.3	75	0.53	28.4	7.8
587	92	0.02	20.3	8.2	0	—	—	—

The first column shows the group number. The two major subdivisions represent all the long-range predictions and those marked by the group as very reliable (≥0.8). Within each subdivision, the first column shows the number of correct predictions, the second the prediction accuracy, and the third and fourth columns show the average residue–residue separation in sequence and the standard deviation, respectively.

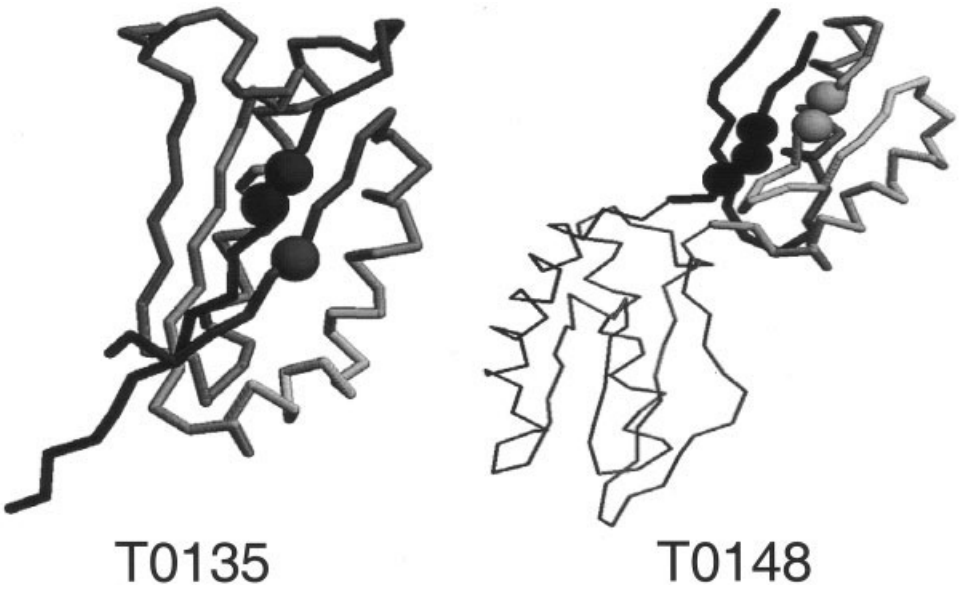

Fig. 13. Residue–residue contact predictions. Backbone representation of the targets T0135 and T0148. Spheres indicate residues predicted to be in contact by the CIRB group (397).

useful. For instance, they predicted a β-strand inserted in the middle of a β-sheet >60 residues apart for targets T0135 and T0148 (Fig. 13).

DISCUSSION

Coordinate Predictions

The accuracy of coordinate predictions for proteins adopting new folds is impressive. That the topology of complicated structures can somehow be teased out despite no prior knowledge of anything resembling the correct fold suggests that great progress has been made toward a general understanding of protein structures and how they fold and evolve. However, some proteins are still not predictable (e.g., T0146_1 and 3, T0172_2, T0187_1) and the overall performance of selecting the best of five models is still poor. There is thus still room for improvement.

The overlap between fold recognition and de novo/ab initio predictions suggests that we are witnessing a general convergence of structure prediction methods. The resemblance between traditional fold recognition and the successful fragment-based structure prediction approaches is clear, and the ability of the fragment assembly approaches to fill the gap between new folds and remote homologues is encouraging. A tendency in previous CASPs was for analogous fold recognition methods to identify correct folds but to fail to align the sequence to the template correctly. De novo prediction methods get around this problem by avoiding any explicit alignment to a template. However, it is also clear that progress has been made in analogous fold recognition: many predictions of equal or even better quality were also made using templates (see Tables II and III).

For fragment-based methods, two main points are clear. The first is that choice of the template library affects performance of the method; apparently similar methods give different results, probably only owing to differences in library construction. The second is that fragment assembly still requires a good deal of human intervention. Servers, when they do well, tend still to find local structures, as did at least one other method where most effort to

date had been concentrated on fragment identification instead of assembly (Kevin Karplus, personal communication).

Among the most impressive successful predictions were those made by the Baker group (002) for small domains appended (T0173_2) or inserted (T0186_3) into others recognizable by sequence comparison or fold recognition methods. These show how de novo predictions methods can indeed fill important gaps in knowledge that inevitably are created when structures are generated by comparative modeling.[23] They are also a subset of a wider phenomenon whereby separate 3D structures are known or predicted for connected and/or interacting domains. A major problem in such instances is to identify the correct relative domain orientation. Insights might come from the protein-docking community (CAPRI[24]) or possibly from examples, such as the Baker prediction for T0129, where orientations were predicted with remarkable accuracy.

Secondary Structure Predictions

At first glance, it might seem that the higher SOV scores seen in CASP5, compared to CASP4, represent a real improvement over the past 2 years. However, inspection shows that this is probably due to a higher proportion of targets homologous to a known structure and/or the fact that some groups extracted secondary structures from comparative models. We may be reaching the limits of secondary structure prediction accuracy as previously suggested[18,25,26] where results can only be improved by physically building and somehow refining 3D structures. However, we do not think that this process is necessarily measured when structures are constructed by comparative modeling techniques.

Secondary structure prediction is a mature field, and we think there are several reasons why it can no longer be effectively evaluated in CASP. First, the small differences in the accuracies observed between sequence-related and-unrelated targets, as well as the fact that several fully automated servers are among the best predictors, suggest that manual intervention does not improve secondary structure prediction accuracy. Moreover, we think that few predictors now focus explicitly on the problem of predicting secondary structures in this way.

A second reason is that the small number of targets considered in CASP makes it difficult to discern differences in group performance. It has been recently shown that, because of large standard deviations, >200 predictions are needed to distinguish a 3% difference in prediction accuracy.[27,28] It has also been argued that some results obtained in CASP experiments might lack statistical validity,[29] particularly when it comes to the problem of ranking groups that tend to perform similarly.

Finally, there is an ongoing effort to evaluate secondary structure (and other) predictions continuously with as many targets as possible to afford the most accurate assessment and ranking possible.[28] We would suggest that the CASP6 assessors leave evaluation to these efforts and would encourage a presentation of these results at every subsequent CASP so that the community can continue to monitor progress in this field. However, we would encourage these efforts to allow for other ideas, such as the inclusion of more complete alphabets for secondary structure types, to be tested.

Residue–Residue Contacts

Only one group was fully evaluated in this category during CASP4, making it difficult to chart progress. Although recent publications[30] have reported overall prediction accuracies as high as 60%, we have not seen such accuracies here, and the results suggest that there has been at best a very limited improvement for de novo methods. Better accuracies have been obtained by using comparative models to derive contacts, but we question whether these predictions have any meaning beyond an unconventional means for assessing comparative model quality. We would not encourage predictors in the future to do this, particularly because it can obfuscate attempts to measure progress for de novo methods.

New contact prediction methods that make use of sequence profiles and fragment templates have been used for the first time in CASP5[31] with accuracies comparable to or even higher than the best classical de novo approaches. These hybrid methods aim to use contact maps as a two-dimensional representation of 3D features and show good potential for future development in this field.[32]

The current state-of-the-art in contact prediction is still a long way from the goal of being an intermediate between secondary structure and full 3D coordinate predictions. As yet, methods are not sufficiently accurate to be used to build a 3D model from scratch, such as input for a constraint satisfaction procedure akin to that used during NMR structure determination. However, they can produce valuable information that might be applicable to discriminating between several potential models generated by other means (e.g., fragment assembly).

Final Thoughts for CASP6 and Beyond

Proteins adopting new folds are increasingly rare, and indeed, even those classified during CASP5 often showed partial similarities to other structures, sometimes even with the best predictions coming from templates covering most of the structure. This begs the question as to whether the impressive progress seen is in part due to an increase in data on which to base empirical structure prediction methods. It is somewhat ironic that we approach a solution to the problem of de novo protein structure prediction just when structural biology promises to provide most of the answers.

On the basis of our assessment in this category, we would suggest that future efforts focus on the remaining difficult tasks of fragment assembly, model and fragment selection, and template refinement, or indeed on other areas of biology (structural or otherwise) where there is a growing need for computational expertise to solve complex problems.

ACKNOWLEDGMENTS

We thank John Moult, Tim Hubbard, Nick Grishin, Lisa Kinch, Anna Tramontano, Veronica Morea, Krzysztof Fide-

lis, Adam Zemla, and Ceslovas Venclovas for an exciting 3 months. We are grateful to the many structural biologists who provided valuable coordinate data before publication. It was an honor to have been chosen to be CASP assessors. We greatly appreciate the hard work that went into the predictions and sincerely hope that we have done justice to it during our assessment.

REFERENCES

1. Defay T, Cohen FE. Evaluation of current techniques for ab initio protein structure prediction. Proteins 1995;23:431–445.
2. Lesk AM. CASP2: report on ab initio predictions. Proteins 1997; Suppl 1:151–166.
3. Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. Proteins 1999;Suppl 3:149–170.
4. Lesk AM, Lo Conte L, Hubbard TJ. Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, and interresidue contacts. Proteins 2001; Suppl 5:98–118.
5. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 1998;281:565–577.
6. 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.
7. Bonneau R, Strauss CE, Rohl CA, Chivian D, Bradley P, Malmstrom L, Robertson T, Baker D. De novo prediction of three-dimensional structures for major protein families. J Mol Biol 2002;322:65–78.
8. Fukushima K, Kikuchi J, Koshiba S, Kigawa T, Kuroda Y, Yokoyama S. Solution structure of the DFF-C domain of DFF45/ICAD. A structural basis for the regulation of apoptotic DNA fragmentation. J Mol Biol 2002;321:317–327.
9. MacLachlan AD. Gene duplications in the structural evolution of chymotrypsin. J Mol Biol 1979;128:49–79.
10. Hubbard TJ. RMS/coverage graphs: a qualitative method for comparing three-dimensional protein structure predictions. Proteins 1999;Suppl 3:15–21.
11. 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;34:220–223.
12. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins 1999;Suppl 3:22–29.
13. Sayle RA, Milner-White EJ. RASMOL: biomolecular graphics for all. Trends Biochem Sci 1995;20:374.
14. Russell RB, Barton GJ. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins 1992;14:309–323.
15. Kinch LN, Qi Y, Hubbard T, Grishin NV. Target classification. Proteins 2003;Suppl 6:340—351.
16. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
17. 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.
18. Rost B, Sander C, Schneider R. Redefining the goals of protein secondary structure prediction. J Mol Biol 1994;235:13–26.
19. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
20. Tramontano A, Morea V, Assessment of homology based predictions in CASP5. Proteins 2003;Suppl 6:352–368.
21. Baldi P, Brunak S, Frasconi P, Soda G, Pollastri G. Exploiting the past and the future in protein secondary structure prediction. Bioinformatics 1999;15:937–946.
22. Karplus K, Karchin R, Barrett C, Tu S, Cline M, Diekhans M, Grate L, Casper J, Hughey R. What is the value added by human intervention in protein structure prediction? Proteins 2001;Suppl 5:86–91.
23. Vitkup D, Melamud E, Moult J, Sander C. Completeness in structural genomics. Nat Struct Biol 2001;8:559–566.
24. Vajda S, Vakser IA, Sternberg MJ, Janin J. Modeling of protein interactions in genomes. Proteins 2002;47:444–446.
25. Kabsch W, Sander C. How good are predictions of protein secondary structure? FEBS Lett 1983;155:179–182.
26. Russell RB, Barton GJ. The limits of protein secondary structure prediction accuracy from multiple sequence alignment. J Mol Biol 1993;234:951–957.
27. Rost B, Eyrich VA. EVA: large-scale analysis of secondary structure prediction. Proteins 2001;Suppl 5:192–199.
28. Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Fiser A, Pazos F, Valencia A, Sali A, Rost B. EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics 2001;17:1242–1243.
29. Marti-Renom MA, Madhusudhan MS, Fiser A, Rost B, Sali A. Reliability of assessment of protein structure prediction methods. Structure (Camb) 2002;10:435–440.
30. Pollastri G, Baldi P, Fariselli P, Casadio R. Improved prediction of the number of residue contacts in proteins by recurrent neural networks. Bioinformatics 2001;17 Suppl 1:S234–S242.
31. Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol 2000;301:173–190.
32. Shao Y, Bystroff C. Predicting inter-residue contacts using templates and pathways. Proteins 2003;Suppl 6:497–502.
33. Kraulis PJ. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr 1991;24:946–950.
34. Merritt EA, Bacon DJ. Raster3D: photorealistic molecular graphics. Methods Enzymol 1997;277:505–524.