WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Domain: Prediction

Assessment of predictions submitted for the CASP7 domain prediction category

Michael Tress,[1]* Jianlin Cheng,[2] Pierre Baldi,[3] Keehyoung Joo,[4] Jinwoo Lee,[4] Joo-Hyun Seo,[4,5] Jooyoung Lee,[4] David Baker,[6,7] Dylan Chivian,[8] David Kim,[6] and Iakes Ezkurdia[1]

[1] Structural and Biological Computation Programme, Spanish National Cancer Research Centre, Madrid, Spain

[2] School of Electrical Engineering and Computer Science, University of Central Florida, Orlando, Florida

[3] Institute for Genomics and Bioinformatics, School of Information and Computer Sciences, University of California, Irvine, California

[4] School of Computational Sciences, Korea Institute for Advanced Study, Seoul 130-722, Korea

[5] School of Chemical and Biological Engineering, Seoul National University, Seoul 151-742, Korea

[6] Department of Biochemistry, University of Washington, Seattle, Washington

[7] Howard Hughes Medical Institute, Seattle, Washington

[8] Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California

ABSTRACT

This paper details the assessment process and evaluation results for the Critical Assessment of Protein Structure Prediction (CASP7) domain prediction category. Domain predictions were assessed using the Normalized Domain Overlap score introduced in CASP6 and the accuracy of prediction of domain break points. The results of the analysis clearly demonstrate that the best methods are able to make consistently reliable predictions when the target has a structural template, although they are less good when the domain break occurs in a region not covered by a template. The conditions of the experiment meant that it was impossible to draw any conclusions about domain prediction for free modeling targets and it was also difficult to draw many distinctions between the best groups. Two thirds of the targets submitted were single domains and hence regarded as easy to predict. Even those targets defined as having multiple domains always had at least one domain with a similar template structure.

Proteins 2007; 69(Suppl 8):137–151.

Key words: domain boundaries; domain overlap; evaluation; ab initio domain prediction; template-based domain prediction; proteins.

INTRODUCTION

Domain identification is an essential early step in many protein analyses. Dividing a protein into structural or functional subunits is usually the first step in protein structure prediction and can be an important part of functional and structural experiments.[1] Domain boundary prediction is also a vital part of the target selection process in structural genomics.[2]

The task of predicting domain boundaries is far from simple, particularly when there is no structural template on which to base the prediction. Where there is no structural information, predictions are based on multiple alignments along with 1D features such as amino acid frequencies and predicted secondary structure and solvent accessibility, often with the aid of neural networks.

Even when there are structural templates domain prediction can still be difficult. The largest potential pitfall associated with

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat.
The authors state no conflict of interest.
Grant sponsor: BioSapiens Network of Excellence; Grant number: LSHC-CT-2003-505265
*Correspondence to: Michael Tress, Structural and Biological Computation Programme, Spanish National Cancer Research Centre, c./ Melchor Fernandez Almagro, Madrid, Spain.
E-mail: mtress@cnio.es
Received 1 March 2007; Revised 21 May 2007; Accepted 5 June 2007
Published online 6 August 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21675

domain boundary prediction is domain definition itself. Protein domains can be defined in a number of ways, as compactly folded structures with their own hydrophobic core,[3] evolutionarily and functionally distinct entities[4] or sub-units that can fold independently into stable tertiary structures.[5] Often these three subdefinitions coincide, making domain delimitation rather simple. Domain boundary definitions become more difficult and subjective when these roles are less clear or when one of the definitions is at odds with the others. The specific training of domain prediction methods may mean that they are more inclined to predict one or other of these definitions.

However, the format of the CASP7 experiment meant that the prediction of independently folded stable structures could not be assessed. All the CASP7 targets were submitted to the organizers because their structure was in the process of being solved; the submitted target sequences were identical (give or take the odd terminal residue) to the sequences of the final solved structures.

There were further problems that predictors and assessors faced (both in this CASP and in CASP6) relating to the lack of suitable targets. While the number of noncancelled targets in the whole CASP experiment rose by 50% in CASP7 to almost 100 targets, only multidomain targets are suitable for domain prediction evaluation. There were only 17 such targets in CASP6[6] and while the numbers almost doubled in CASP7, there will still only 32 multiple domain targets. In addition there were no targets where the domain prediction could be considered to be completely template-free—there was at least one domain in each of the multidomain targets that had a structural template.

Targets were assigned domains for the structure evaluation by the assessors as detailed elsewhere in this issue.[7] However, it should be borne in mind that the domain assignments used in the assessment of structure prediction were not always the same as those used for the domain prediction assessment. The domain prediction assessors changed domain definitions for several targets and allowed alternative domain definitions for a number of targets.

Here we present the assessment of the CASP7 domain prediction category. We assessed predictions using two separate scoring schemes and were able to show that a number of groups were able to make better predictions than the average. There were statistically significant differences between the best groups and the rest of the predictors. The results of the assessment also allowed us to propose strategies for the analysis of domain prediction in the future.

METHODS

We analyzed the predictions with two separate scoring schemes. The first was the Normalized Domain Overlap scoring system that was introduced in the CASP6 domain assessment.[6] The advantage of this scoring scheme is that it reduces the scoring of the prediction to a single normalized score and it penalizes both under prediction and over prediction of domains. It has one drawback in that it does not explicitly recognize correct prediction of interdomain linkers.

The other score used is a measure of accuracy of domain boundary prediction.

Table I
The Domain Overlaps for a Prediction for Target TO381[a]

	Linkers (77–89)	D1 (17–76)	D2 (90–265)	Score
linker (157–174)	0	0	18	—
d1 (1–156)	13	60	67	−6
d2 (175–265)	0	0	91	91
Score	—	60	6	75.5

[a]Both the prediction and target have two domains and a short linker. Each white cell contains the number of overlapping residues between the predicted domain and the actual domain definition. The predicted domain numbers are preceded by lower case identifiers (rows), the defined domains by upper case identifiers (columns). The calculation process is explained in the text.

Normalized domain overlap

Normalized domain overlap was introduced in CASP6 and can be followed in more detail in the CASP6 assessment paper.[6] It is calculated as follows: Predicted and official domains are compared and the number of residues that overlap between predicted domains and linkers, and defined domains and linkers are summed for each domain. For example a target structure that was split into two domains and a linker in both target and prediction would require a total of nine overlaps to be calculated.

In the example in Table I (a prediction for target T0381) there are nine possible overlaps arising from the domain definition and the prediction—both have two domains and a linker. If there is more than one linker, leader, or trailer in the assignment or prediction they are treated together.

The total overlap score can be calculated from the matrix. For each column and row (excepting the linkers) each of the smaller scores are subtracted from the largest score. For example in column D2, 67 and 18 are subtracted from 91. No score is computed for the combined linkers in either the definition or the prediction. The total overlap score is the sum of all the column and row subtotals from the defined and predicted domains, divided by 2. The final score is shown in bold in the bottom right-hand corner of Table I.

The normalized domain overlap (NDO) score is simply the total score divided by the number of residues in the defined domains. When there is more than one defined domain, scores are calculated for each of the domain definitions and the best score is taken as predictor score for that target.

DOI 10.1002/prot

Domain boundary distance score

Simple domain boundary scores reward all predictions that are within a certain cut-off of the correct domain boundary. Under our scoring scheme all predictions within 8 residues of the correct boundary will score, but predictions that are closer to the correct domain boundary would score more.

All distances between the predicted and correct domain boundaries are calculated. If the domain boundary has a linker, the whole linker is regarded as the domain boundary. In a fashion similar to the GDT scoring scheme employed in the structure prediction category, predictions are given one point for being within 1 residue of each correct boundary, another point if they are within two residues, a further point if they are within three, and so on up to eight residues. A prediction two residues away from the correct boundary would therefore have 7 points.

The total score for each domain prediction is then calculated as the sum of all predicted boundary scores divided by eight and the total number of domain boundaries. The number of domain boundaries comes from either the target or the prediction, whichever is higher. In this way over-prediction is penalized.

Domain assignment

Official domain definitions for the evaluation of structure prediction were made by eye by the assessor for the high accuracy prediction category with input from three other assessors.[7]

The domain definitions used in the domain prediction category were not identical to those in the structural prediction category. The domain prediction category included the three assessor-cancelled targets (T0310, T0343, and T0352) and while the structure prediction category definition was retained for the majority of the targets, for some targets we felt that a different domain definition was more suitable for the domain prediction category.

Domains were assigned for the domain prediction category by visual inspection. Domain boundary decisions were based on structural integrity, on whether each domain had a hydrophobic core, on whether the target had internal symmetry, on evolutionary relationships inferred from the LGA[8] superpositions and sequence alignments, and on the SCOP[9] and CATH[10] domain annotations from those structural templates found by LGA.

A number of targets had no single straightforward domain definition. Where we felt that there was more than one way to split the structure into domains, the target was assigned up to four alternative definitions that were equally valid. In all scoring schemes if a target had multiple domain definitions, scores were calculated for each of the alternative domain definitions and the best score was used in the evaluation.

One drawback of allowing more than one domain definition per target is that if one of the definitions is a single domain, the target effectively becomes a single domain target. Single domain targets are much easier to predict and it is much easier to score full marks with a single domain target. Targets with single domains were not used in the comparisons and for that reason we tried to avoid assigning alternative single domain definitions to official multidomain targets.

Three targets illustrate the domain definition strategy of the assessors. In the structure prediction assessment target T0291 was a single domain protein and target T0292 was split into two domains, even though both proteins belong to the same structural family. While there were good reasons for not splitting T0291 in the structure prediction evaluation, the domain assessors felt that the target was clearly a two-domain protein in the classical sense. We allowed the two-domain definition for both targets, but the single domain alternative was not allowed.

T0299 (Fig. 1) was clearly a two-domain protein, but we allowed four alternative domain definitions. T0299 is formed from two alpha-beta sandwich domains that are joined by two decorative helices. The domains have structural templates, but none of the templates have the decorative helices. Hence it is unclear to which domain the two helices belong. The four alternative domains [Fig. 1(a–d)] allow four possible arrangements of the helices and domains, but penalize predictors if they predict the domain boundaries in any other part of the decoration.

It is important only to define linkers where it is clear that the linking residues do not belong to one or other of the domains. Target T0289 (Fig. 2) illustrates this very well. It is clearly a two-domain protein. Between the two domains are two long anti-parallel beta-strands that also interact with the N-terminal. These two strands could be regarded as linkers—they are present in very few of the structurally similar parent structures—but they quite clearly interact with the surface of domain 1. If they are defined as linkers in an alternative definition, the prediction of domain boundaries becomes an exercise in structural template detection because then it does not matter where the linker is cut. For example the prediction in Figure 2(b) would score full marks, even though it is clearly not quite correct. In this case the linking strands were defined as being part of domain 1 and there was no alternative definition.

RESULTS

A total of 26 groups made predictions in CASP7. Twelve groups were "human," though three of these groups predicted very few targets and their results do not appear in this paper. Fourteen groups made server pre-

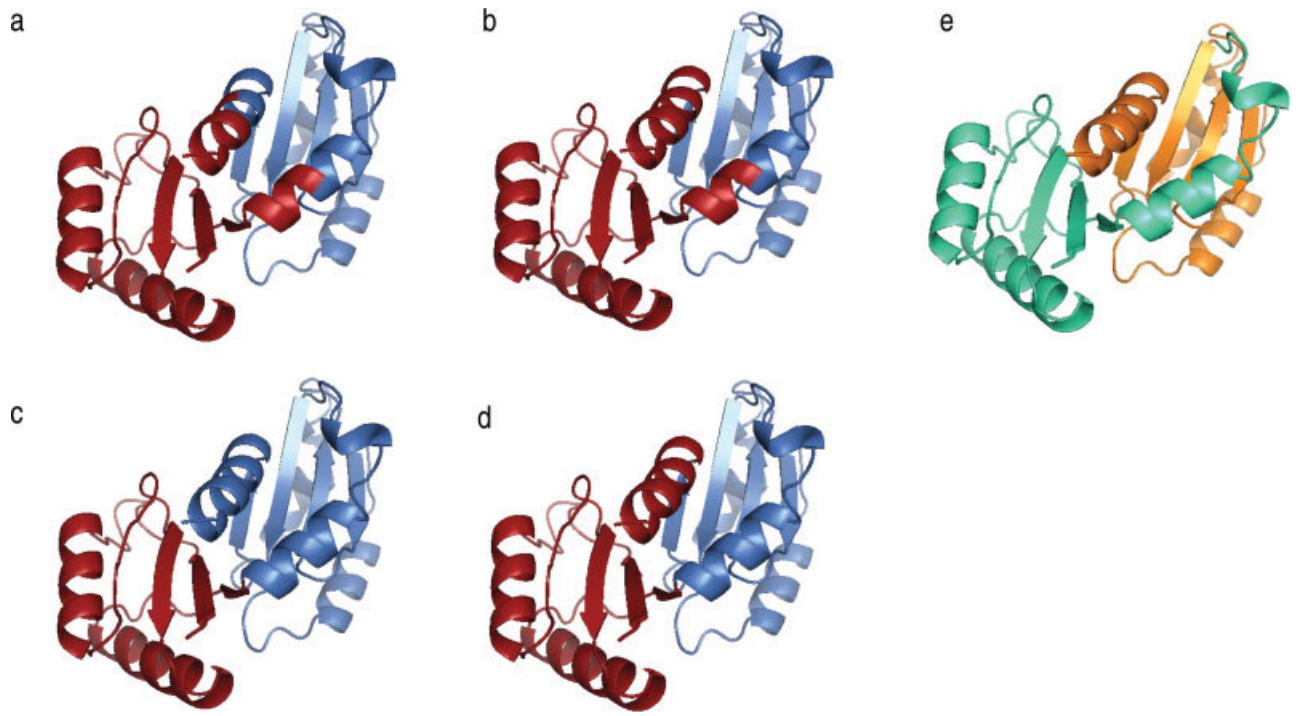

Figure 1

*Alternative domain definitions for T0299. Here we show the four alternative (and equally valid) domain definitions for target T0299 (**a**–**d**). Domain 1 is in blue, domain 2 in red and the differences in the definitions are in the linking helices. Part (**e**) shows a good prediction for this domain, but one in which the prediction for domain 2 is encroaching into the official domain 1. Pymol was used to generate Figures 1, 2, 6 and 7. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]*

dictions. There were 98 structures in the assessment, 30 had alternative domain definitions (see supplementary Table I for the full details), and 31 targets were considered to be multiple domain for the purposes of the assessment; that is none of the domains or alternatives were single. The equivalent figures in CASP6 were 63 targets, 20 with alternatives and 17 multiple domain targets.

Initial comparisons of the mean NDO scores for all 98 targets confirmed our suspicions that most of the targets were not adding anything to the evaluation. For the vast majority of the single domain targets virtually all the predictors had high NDO scores (Fig. 3). For 10 single domain targets every predictor has perfect scores.

It was clear that the evaluation should be based on a subset of targets and the most obvious subset to consider would be the multiple domain targets (as in CASP6). We did also carry out an assessment based on what we considered were the "hard targets," a subset of 24 targets for which the mean NDO scores were less than 80%. This group included one single domain protein, T0287, a free modeling target with no sequence homologs that seemed to cause predictors some difficulties, but excluded seven easy multiple domain targets. The subset of hard targets was chosen because a number of the multiple domain targets in this issue of CASP seemed particularly easy and we felt they were only contributing noise to the comparison. The predictions for the hard subset were subjected to the same assessments as those of the multiple domain subset. There was little difference between the two sets of results. The results for the "hard" subset can be seen in supplementary Figure 1.

Nine groups have slightly better mean NDO scores than the others over all the targets. The same 9 groups stand out when mean NDO scores are compared over the subset of multiple domain targets [Fig. 4(a)], but the differences are far more pronounced. The groups DP105, DP136, DP497, DP556, DP581, and DP722 have slightly better predictions than groups DP091, DP229, and DP686. These results were repeated when the comparisons were made with the *Z*-scores of each prediction in order to account for possible differences in the difficulties of each target (see supplementary Fig. 2).

Domain boundary distance scores

Predictions were also evaluated by accuracy of domain boundary placement for multiple domain targets. Here a

DOI 10.1002/prot

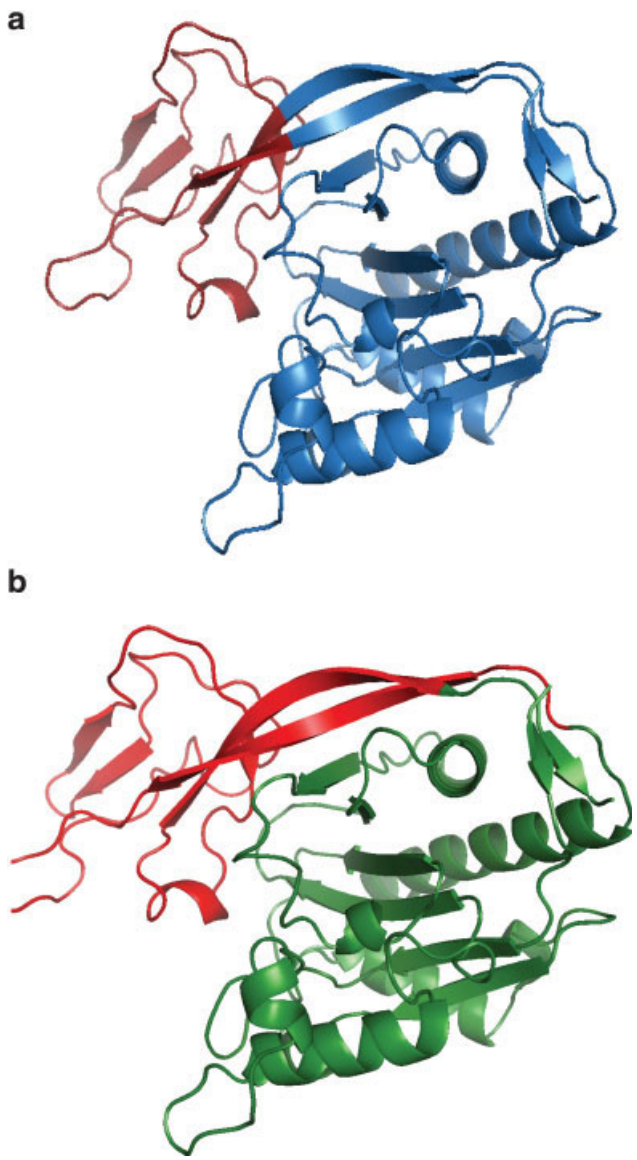

Figure 2

*(**a**) Domain definition for T0289. Target T0289 has just a single domain definition. Domain 1 is in blue, domain 2 in red. The long comb-over sheet at the top if the structure does not have a template, but the sheet was defined as part of domain 1 as it in contact with the domain 1 for much of its length. In (**b**) we show one of the better predictions, but where the lack of a template to model the comb-over means that most of the strand is predicted as being part of domain 2. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]*

similar pattern emerges. Figure 4(b) shows the domain boundary prediction scores for the subset of multiple domain targets and the same six groups (DP722, DP581, DP105, DP497, DP136, and DP556) have slightly better mean domain boundary distance scores than groups DP091, DP229 and DP686.

Sensitivity and accuracy of domain boundary prediction

The domain boundary distance score combines the accuracy and sensitivity of domain boundary prediction into a single score. However, it is also useful to look at domain boundary prediction from the separate perspectives of accuracy (fraction of predicted boundaries that are correct) and sensitivity (fraction of "true" boundaries that are correctly predicted). Viewing both measures is necessary because it allows one to avoid placing too much emphasis on over-predicting boundaries to obtain completeness at the expense of accuracy, or under-predicting boundaries to obtain accuracy without sufficiently obtaining the domain edges. Since exact agreement between the residues predicted as the boundary and the "true" boundary is unusual, the measures should be viewed over a range of distances from the true boundary. Sensitivity and accuracy for the best performing predictors is shown in Figure 5. It is clear that there are differences between predictors, with some groups erring towards over-predicting boundaries and others towards under-predicting boundaries.

Statistical comparisons

Groups were compared head-to-head over common subsets of predicted targets. We carried out paired *t*-tests between each pair of groups over common targets using both the NDO and domain boundary prediction scores

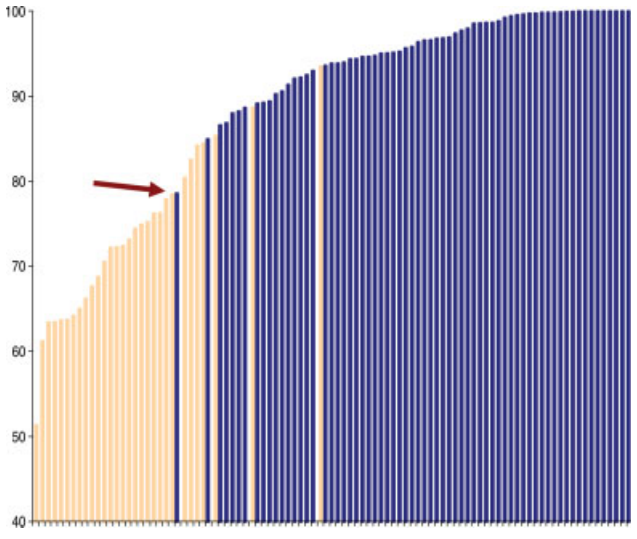

Figure 3

Mean NDO scores for the targets. Single domain targets are in dark blue, multiple domain targets are in light orange—it is clear that the multiple domain targets are much more difficult to predict. The arrow shows where we split the targets into the hard and easy subsets. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

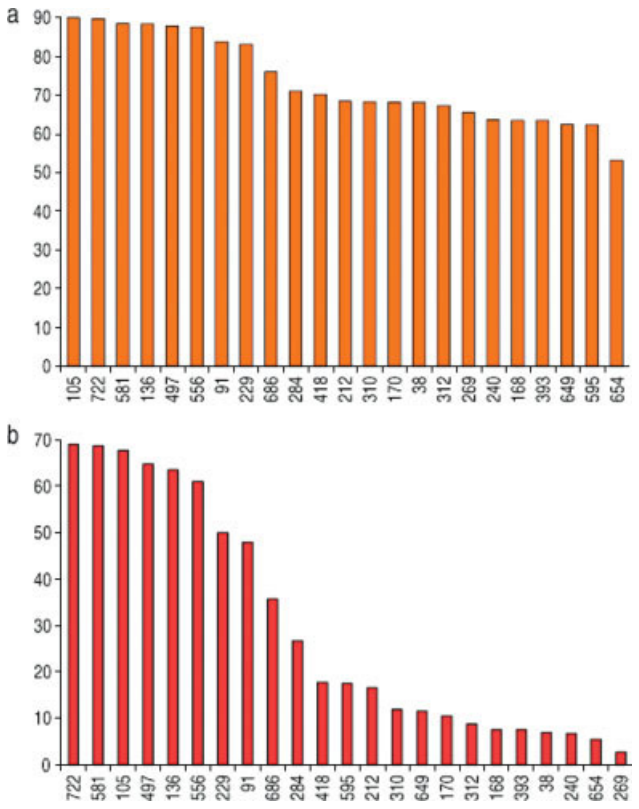

Figure 4

*Mean NDO and domain boundary scores for each group. In (**a**) the mean NDO scores for each group over the subset of multiple domain targets. In (**b**) the mean domain boundary scores for each group over the subset of multiple domain targets. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]*

and the *P* values from these comparisons are shown in Table II. The results of the domain boundary score comparison shows that five groups (DP136, DP497, DP556, DP581, and DP722) have significantly better scores than the other groups. Group DP105 is an exception—they predicted just less than half the multiple domain targets—so they had fewer common targets and it was difficult for differences to be significant. Predictors DP091 and DP229 have significantly better domain boundary scores than all but the six groups already mentioned and group DP686. The results from the comparison of NDO scores are similar, but the distinctions between the groups are a bit more blurred.

Best scoring groups

The groups with the best predictions can be divided into two groups: those methods that are built to make template-based domain predictions and those groups that can make hybrid template-based/ab initio predictions.

The purely template-based methods were the Lee Group (DP556), Andante (DP105), and Ginzu (DP581). The groups using hybrid methods were FoldPro/DomPro (DP136), Ma-Opus-Dom (DP229 and DP091), Rosetta-DOM (DP497), and DP_Hybrid (DP722). DP136, DP581, DP497, and DP229 were all server groups. Five of the methods are described in more detail by their authors in this paper.

It was no surprise that the top predictors were template-based or hybrid methods because all the free modeling targets (the more difficult targets) were single domains or part of multiple domains in which the prediction could be made by subtracting the template-based domain(s) of the target. A few template-based targets, such as T0321 or T0356, were more difficult to predict because templates were harder to find. There were some outstanding predictions for these targets such as the predictions from DP581 and DP229 for target T0301 (Fig. 6) and from DP310 for T0356 (Fig. 7). Single do-

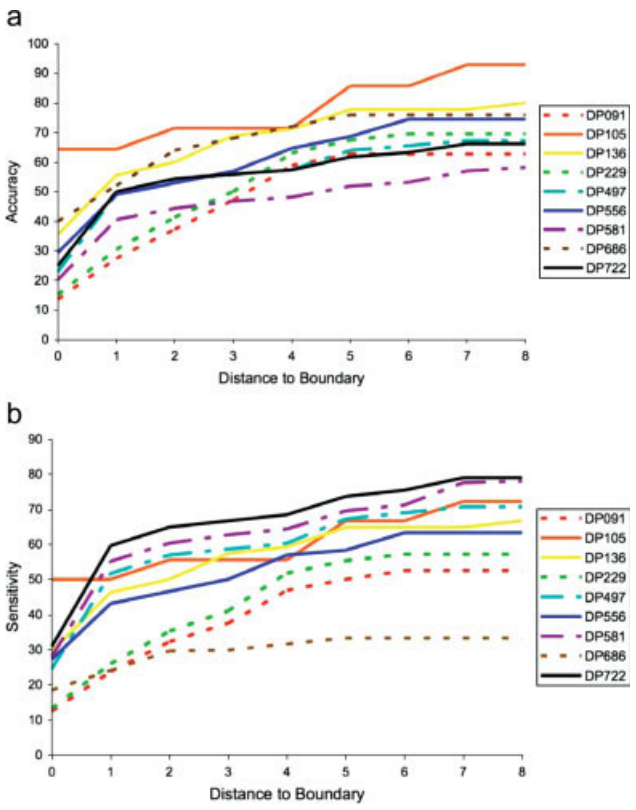

Figure 5

*Accuracy and sensitivity of multi-domain target predictions. Accuracy of domain prediction is shown in (**a**), sensitivity shown in (**b**) predicting groups are color-coded. Accuracy and sensitivity are measured at distance tolerance intervals of one to eight residues from the "true" boundary. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]*

DOI 10.1002/prot

Table II
P *Values for the Paired* t*-tests for NOD Scores and Domain Boundary Scores*[a]

	DP722	DP136	DP556	DP105	DP497	DP581	DP091	DP229	DP686	DP284	DP418	DP310	DP038	DP170	DP212	DP312	DP269	DP240	DP168	DP393	DP649	DP595	DP654
DP722		0.42	0.19	0.12	0.05	0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DP136	0.61		0.82	0.22	0.86	0.83	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DP556	0.50	0.82		0.97	0.62	0.95	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DP105	0.58	0.71	0.67		0.12	0.58	0.45	0.45	0.25	0.09	0.01	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.00
DP497	0.10	0.86	0.98	0.58		0.93	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DP581	0.05	0.47	0.20	0.12	0.06		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DP091	0.06	0.03	0.07	0.79	0.21	0.27		0.79	0.10	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DP229	0.01	0.01	0.01	0.73	0.06	0.04	0.38		0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DP686	0.00	0.00	0.00	0.66	0.00	0.01	0.02	0.04		0.29	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00
DP284	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.08		0.30	0.13	0.02	0.04	0.23	0.03	0.00	0.01	0.01	0.01	0.21	0.29	0.02
DP418	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.07	0.72		0.29	0.11	0.14	0.18	0.04	0.02	0.05	0.12	0.12	0.16	0.82	0.06
DP310	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.98	0.45		0.19	0.90	0.38	0.56	0.05	0.28	0.43	0.43	0.67	0.62	0.28
DP038	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.87	0.42	1.00		0.63	0.14	0.67	0.37	0.96	0.77	0.77	0.56	0.24	0.64
DP170	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.01	0.91	0.30	0.86	0.70		0.18	0.23	0.12	0.43	0.56	0.56	0.78	0.56	0.34
DP212	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.04	0.92	0.11	0.69	0.52	0.62		0.05	0.03	0.09	0.16	0.16	0.20	0.95	0.08
DP312	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.01	0.88	0.26	0.83	0.88	0.56	0.50		0.20	0.68	0.84	0.84	0.83	0.32	0.53
DP269	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.59	0.09	0.51	0.63	0.24	0.24	0.60		0.38	0.28	0.28	0.12	0.05	0.27
DP240	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.30	0.02	0.28	0.41	0.06	0.08	0.19	0.46		0.76	0.76	0.49	0.24	0.83
DP168	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.04	0.09	0.10	0.13	0.09	0.26	0.31	0.58			0.44	0.18	0.71
DP393	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.04	0.09	0.10	0.13	0.09	0.26	0.31	0.58			0.44	0.18	0.71
DP649	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.41	0.01	0.08	0.37	0.07	0.03	0.17	0.17	0.87	0.65	0.65			0.35
DP595	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.26	0.01	0.03	0.16	0.01	0.01	0.11	0.09	0.49	0.85	0.85	0.33		0.07
DP654	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	

[a]The paired *t*-test *P* values for the NDO scores shown in the bottom left half of the table. The paired *t*-test *P* values for the domain boundary scores are shown top right. Significant differences between groups over common sets of predictions are shown with a white background; where there are no significant differences the cells are shaded in dark grey with white font. The lighter shades of grey show where the *P* values either side of the limit are close to the cut-off for significance (0.05).

main predictions for the free modeling target T0296 were also impressive given that the protein has more than 500 residues and that no structure prediction group detected a structure.

Group reports

Domain prediction group 136 (DOMpro)

Inspired by the success of hybrid prediction approaches[11] in the CASP6 experiment, we developed a hybrid domain predictor for CASP7, integrating template-based structure prediction, domain parsing, and ab initio domain prediction.

First, the FOLDpro module of the SCRATCH suite[12] is used to extract pairwise similarity features for the target and all the templates in a fold recognition library of about 10,000 proteins. The similarity features are fed into a support vector machine (SVM)[13] to evaluate the structural relevance of each template and rank the templates accordingly. PSI-BLAST[14] searches of the whole PDB[15] are included to improve the model quality for easy targets.

If the SVM score of the top-ranked template is greater than a threshold, FOLDpro uses PSI-BLAST and COACH[16] to generate alignments for easy targets and hard targets respectively. The top-ranked template-target alignments are fed into Modeller[17] to generate model structures. Finally DOMpro uses PDP[18] to parse the model into domains. If the assigned domains do not cover the entire sequence, DOMpro assigns uncovered regions to the closest adjacent domain.

If the SVM score is less than the threshold, DOMpro proceeds with ab initio domain prediction.[19] DOMpro tries to predict whether or not a residue is in a boundary region (at most 20 residues away from a real domain boundary) using one-dimensional recurrent neural networks[20] in conjunction with profiles, along with predicted secondary structure and solvent accessibility. The protein is cut into domains according to those positions predicted to be in boundary regions.

The evaluation of DOMpro in CASP7 provides valuable insights. For template-based domain prediction, our experiments show that domain-parsing tools, such as PDP, are quite effective at decomposing predicted models into domains. They are robust to model inaccuracies, as long as the coarse topology of the structure is predicted correctly. Figure 8(a) shows a perfect example (T0323), where DOMpro predicts the exact domain boundaries, despite the 2.9 Å RMSD between the model and the true structure.

However, as previous observations have shown,[22] PDP may disregard secondary structure elements in cutting proteins, as shown in Figure 8(b) (T0318) and may encounter difficulties with complex and ambiguous domains. Domain ambiguity is common—there are alternative domain definitions for about one third of CASP7 targets. In ambiguous cases PDP tends to "over cut" into multiple domains.[22]

Our results show that combining template-based 3D structure prediction and domain parsing techniques is an

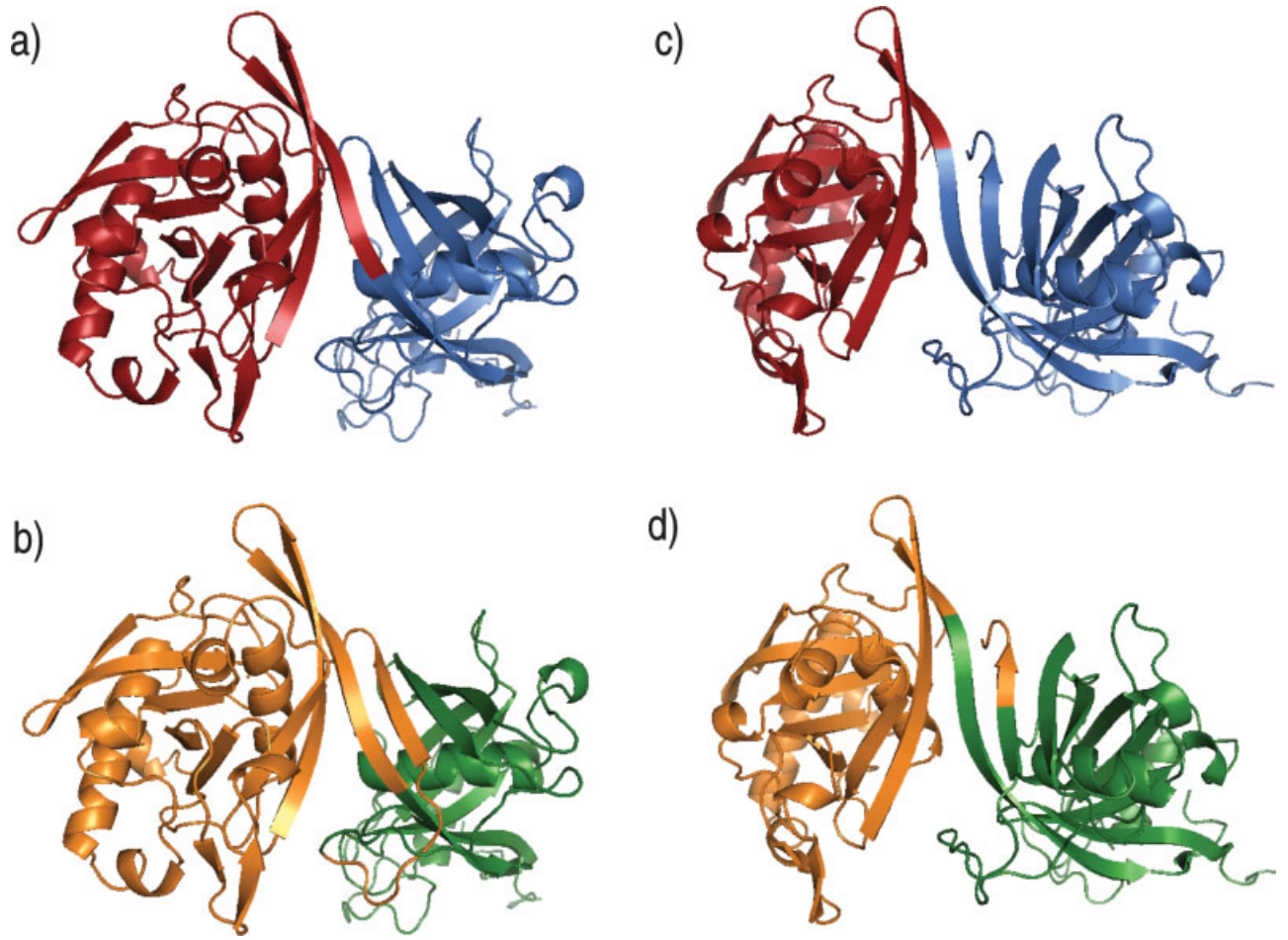

Figure 6

*Two outstanding predictions for T0301. In (**a**) one of the two alternative definitions of target T0301. In (**c**) the prediction from group DP581. T0301 is a target that should have been difficult to predict. Note that this is a non-continuous domain and that while one of the domain boundaries is predicted perfectly the other boundary overlaps into domain 1. NDO score of 90. In (**b**) the other alternative definition for target T0301. In (**d**) the prediction from group DP229 corresponding to this definition. Again note one of the domain boundaries is predicted perfectly, while the other boundary overlaps into domain 1. NDO score of 96. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]*

effective approach where overall coarse topology can be predicted.

For *ab initio* domain prediction, the neural network approach using profiles and predicted structural features can make useful predictions. DOMpro correctly predicts ab initio the large T0296 target as a single domain, although it fails to predict the two domains of T0347. Overall, in CASP7 DOMpro does well on the majority of the ab initio targets. However, the performance is likely to be an overestimate since there are only 12 relatively small *ab initio* targets.

As previous research[19,23,24] and CASP6 have shown, the accuracy of *ab initio* domain prediction methods for multidomain proteins is still low, significantly lower than for template-based methods. Thus, it is important to further improve the accuracy of *ab initio* domain prediction. Since protein domains are largely shaped by gene recombination events, such as gene fusion, fission, domain swapping, and exon exchange (see T0379 for a likely example of fusion), leveraging these evolutionary gene recombination signals, embedded in the multiple sequence alignment of a protein or even in its gene structure, may help improve ab initio domain prediction.

Domain prediction group 556–LEE

We present a domain prediction procedure based on predicted 3D models. The procedure includes five steps: fold recognition, multiple sequence alignment, 3D modeling, prediction assessment, and domain parsing.

The first step is to prepare several sets of templates by fold recognition. To collect fold candidates, we considered top scoring templates from the meta-server 3D-Jury[25] and an in-house method called FoldFinder (Joo

DOI 10.1002/prot

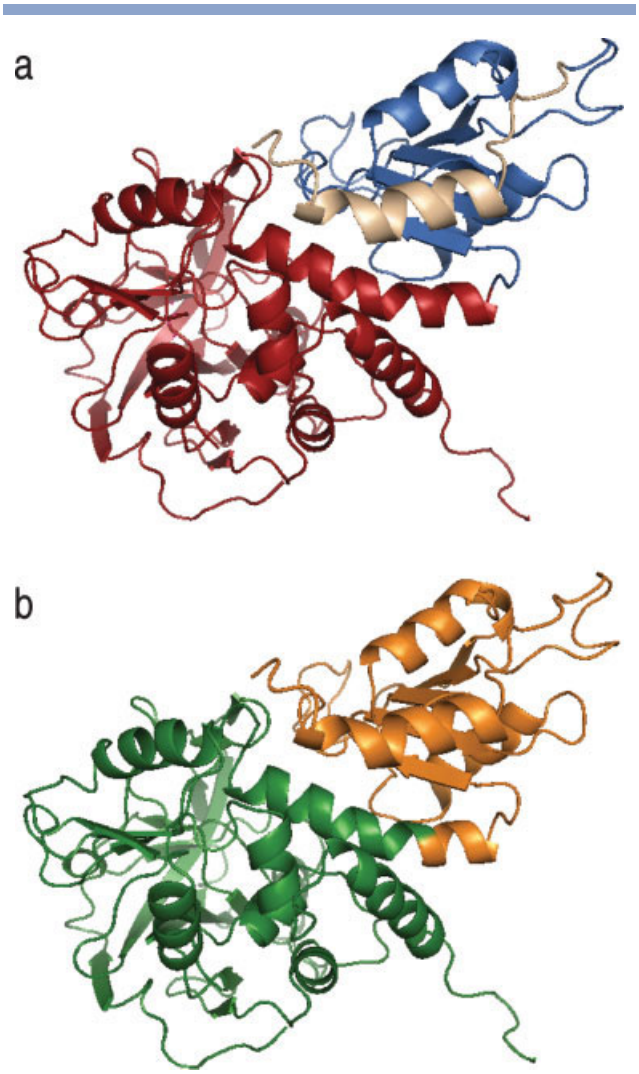

Figure 7

*Outstanding prediction for T0356. In (**a**) one of the three alternative definitions of target T0356. In (**b**) the prediction from group DP310. This prediction is especially good because T0356 is the nearest thing that CASP7 has to a multiple domain free modeling target. There was a structural parent for one of the three structure prediction category domains, but only 9 predictors found the template.*

et al., in preparation). FoldFinder is a profile-profile alignment method utilizing predicted secondary structure. We used a fold database of 17,930 protein chains from PISCES[26] at the 99% sequence identity level. With these templates we performed structural clustering, which typically led to 2 to 3 sets.

The second step is to generate multiple sequence/structure alignments for each template set with MSACSA (Joo *et al.*, submitted for publication). Since model backbones are mainly determined by these alignments, this was the most crucial step in our domain prediction. A unique feature was that we applied a rigorous global optimization method to a score function by using conformational space annealing (CSA,[27]) in contrast to the usual heuristic (progressive) alignment methods popular in the literature. The score function is an in-house consistency-based one. It gives a higher score for alignments that are more consistent with the pair-wise restraint library generated from profile-profile alignments between the query sequence and template sequences and structure-structure alignments between templates. The top scoring alignment from each list was used for 3D modeling.

Modeling was carried out with Modeller[17] after optimizing Modeller energy using CSA. For each multiple alignment, 100 models were generated and all of them were used for the list selection procedure.

We selected the winning list (if one existed) by assessing the average quality of the 100 3D models of each list with an in-house neural-network-based procedure. To select models for domain parsing, we applied a clustering method to find the central models of the 2-3 largest clus-

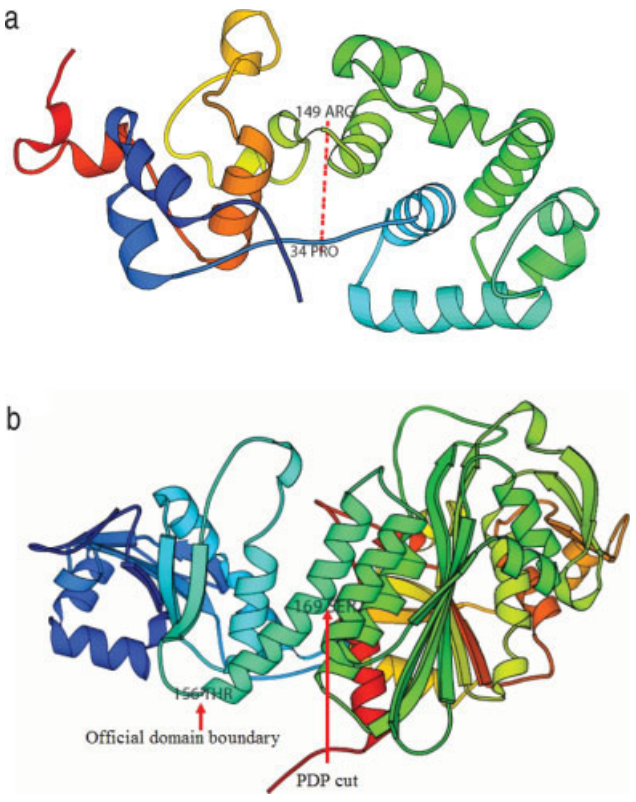

Figure 8

*Two examples of template-based domain prediction. (**a**) A perfect example (T0323). PDP cuts the predicted model into two alpha helical domains (see the virtual line), exactly as in the official domain definition. Domain 1 (left) consists of two non-continuous segments (1–33 and 150–217). Domain 2 spans residues 34 to 149. (**b**) A non-perfect example (T0318). T0318 has two α/β domains, extending from position 1 to 155 and from 156 to 490, respectively. PDP cuts the model in the middle (at residue 169), which is within the long alpha helix. However, a human expert would smartly choose to cut at the end of the helix, avoiding breaking the helix. Figures rendered using Molscript[21]). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]*

ters from the winning list. The best scoring models in terms of the Modeller energy and/or DFIRE[28] energy were also selected. Typically, the central model of the largest cluster served as model 1. When there were competing lists, we used all of them to select 5 models.

The neural network was quite successful at selecting the winning list. The network consists of 5 input nodes, 3 hidden nodes, and 1 output. Inputs are the Modeller energy, DFIRE energy, and the consistencies of a model with 3 predicted properties (secondary structure, solvent accessibility, and hydrophobicity). Details will be published elsewhere. The network was trained so that the output predicts the TM-score of a model. For training and testing, 1,600 models for 6 early-released CASP7 targets were used and 5+1 cross-validations were applied.

PDP and visual inspection were used for domain parsing the 3D models. When the results did not agree, the prediction was chosen by visual inspection. We also consulted the output of PPRODO[29] especially for the hard targets. PPRODO is a sequence-based domain prediction method.

Obviously the quality of the prediction of domains depends on the quality of the models. For our method we found that NDO-scores were greater than 0.9 for a model with TM-scores[30] greater than 0.6. However, there were two exceptions, T0321 and T0341. In the case of T0321 (a hard TBM target), our NDO-score was high, 0.936, although TM-score of our model was quite low, 0.225. The templates used were two domain proteins with similar structures. Although the templates were not good for 3D modeling, the domain boundary was correctly positioned and the success is mainly due to the results of the multiple sequence alignment.

For T0341, although the 3D model quality was relatively high with TM-score 0.84, the NDO-score was only 0.77. PDP defined the first domain as 1-46, 179-259 and, although the position 47 breaks a beta-sheet pair in the middle, we thought it could be acceptable as a proper domain boundary at the time of domain prediction. The official definition of the first domain covers residues 7-74, 179-259 and includes the whole beta-sheet pair.

Domain prediction groups DP497 (RosettaDOM), DP581 (Ginzu), and DP722 (DP_Hybrid)

We investigated the efficacy of several approaches for domain prediction using the Robetta[31] server (http://robetta.org/). These methods have been previously described in CASP6.[11]

"RosettaDOM" is based on the Rosetta[32] method for *de novo* tertiary structure prediction, and produces a 400-member decoy ensemble. Boundary consensus is sought using Taylor's structure-based domain parsing method.[33] Increased frequency of boundaries within a sliding window (smoothed in the same fashion as SnapDRAGON,[34]) is used to assign domain boundaries. Although Rosetta is unlikely to produce atomic-resolution models, it may accurately predict coarse structural features such as domains.

The "Ginzu" method determines boundaries based on the best available information. It begins by scanning for PDB[15] homologs with PSI-BLAST,[14] before searching remaining regions for remote homologs with fold-recognition servers (3D-Jury-A1[25] in CASP7), scanning for domain sequence families with Pfam,[35] and finally looking for conserved sequence blocks within the PSI-BLAST multiple sequence alignment (MSA). Domain boundaries within regions that have PDB homologs are determined using a consensus structure-based domain parser (based on Taylor's method). Boundaries between regions with PDB and/or protein family assignments are determined using sequence edges and low-occupied positions in the MSA and take into account loop regions predicted by PSIPRED.[36]

For targets with PDB homologs the first model submitted for Ginzu came from "Ginzu_HM." This method parses the model built by Robetta as well as PSI-BLAST detectable PDB homologs to make consensus structure-based boundary assignments. Domains are resolved from structure-structure alignments, while attempting to avoid introducing boundaries into helices and strands. "Ginzu_TP," returned as model 2, does not parse the Robetta model directly, rather it consensus parses the parent PDB structure along with PSI-BLAST-detectable PDB homologs. The domain structure of the target is inferred from the K*Sync alignment[37] between the target sequence and the parent PDB. ""Ginzu_SEQ" was only applied to those targets without a detectable homolog or Pfam family (when it was returned as model 1), or to low confidence fold-recognition hits (when it was returned as model 3). It uses sequence information from the MSA to develop a boundary preference function. Note that Ginzu (group DP581) was always represented by model 1 in the assessment.

The "DP_Hybrid" predictor did not participate as a server group in CASP7 but was fully automated. It makes domain boundary predictions by taking advantage of the correspondence between weak signals from RosettaDOM and Ginzu_SEQ. For remote targets Ginzu and RosettaDOM often do not arrive at strongly predicted boundaries, but may suggest several candidate boundaries with confidences below the method thresholds. Agreement between the sequence-based and structure based domain prediction methods increases the confidence of a boundary prediction. DP_Hybrid combines the boundary confidence functions from the two methods and reports boundaries when the combined function is above the threshold, either from a strong prediction by one method or when weaker predictions from different methods are in agreement.

For regions with PSI-BLAST detectable homologs, both DP_Hybrid and RosettaDOM return the Ginzu_HM prediction.

DOI 10.1002/prot

Table III
Single Domain Prediction for Ginzu, RosettaDOM, and DP_Hybrid

	Single domain targets		Multiple domain targets	
Method	NP[a]	Predicted[b] as single	NP[a]	Predicted[b] as single
PSI-BLAST regime[c]				
Ginzu_HM	41	80.5%	21	0%
Ginzu_TP	41	85.4%	21	0%
Remote regime[d]				
Ginzu_HM	14	57.1%	6	0%
Ginzu_TP	14	92.9%	6	16.7%
Ginzu_SEQ	23	95.7%	4	25.0%
RosettaDOM	27	92.6%	6	33.3%
DP_Hybrid	26	92.3%	6	0%

Analysis of over- and under-prediction of single domain targets by the groups Ginzu (DP581), RosettaDOM (DP497), and DP_Hybrid (DP722). Single and multiple domain targets were taken from the structure prediction evaluation as defined by the assessors.[7] Difficulty regimes were defined as those targets for which confident PDB homologs were detected ("PSI-BLAST" regime) or were not detected ("Remote" regime) with the Robetta implementation of PDB-BLAST.
[a]Number of predicted targets attempted by each method within each category.
[b]Percentage of predicted targets ("NP") that were predicted as a single domain.
[c]There were 41 single domain and 21 multiple domain targets in the PSI-BLAST regime.
[d]There were 27 single domain and 6 multiple domain targets in the Remote regime. Ginzu_SEQ was not used in the PSI-BLAST regime and RosettaDOM and DP_Hybird used the Ginzu_HM predictions for these targets. All methods predicted in the Remote regime, but Ginzu_SEQ did not predict targets T0301 and T0321.

An examination of single domain predictions (Table III) reveals that Ginzu_HM is a bit more aggressive than Ginzu_TP in recommending domain boundaries. This is likely a consequence of using the entire model for parsing; long loops that project incorrectly may appear as separate domains to the structure-based parser. Overall, the methods are fairly good at discerning single domain targets, but this result is probably due to the shorter length of most single domain targets. While trends should be viewed with caution given the small sample sizes (in particular for the remote multidomain targets, for which there are only 6 targets), it does appear that the homology-based methods Ginzu_HM and Ginzu_TP are less conservative in predicting single domains than Ginzu_SEQ and RosettaDOM. Encouragingly, the DP_Hybrid method appears to recover correct predictions from the misleading boundaries suggested by Ginzu_HM for the remote single domain targets.

The performance of Ginzu_HM and Ginzu_TP in predicting internal boundaries for multiple domain targets in the PSI-BLAST regime (21 multiple domain targets with 34 internal boundaries) is shown in Figure 9 and reveals an interesting finding. Ginzu_HM (37 predicted boundaries, 29 of which are correct at ±10 residues) does worse than Ginzu_TP (36 predicted boundaries, 32 correct at ±10 residues). This trend is reversed in the Remote regime (only 6 targets with 11 internal boundaries), where Ginzu_HM (11 predicted boundaries, 6 correct at ±10 residues) is better in terms of both specificity and sensitivity than Ginzu_TP (11 predicted boundaries, 4 correct at ±10 residues). It may be that the Robetta homology models aid prediction in the Remote regime, while the model was detrimental in the PSI-BLAST regime, in keeping with our Robetta results in the template-based tertiary structure prediction part of CASP7. Assessors note: these results are borne out by the assessment—Ginzu, DP581, was the only one of the top 9 groups to show a marked improvement when the groups were compared over their best models instead of the first models.

Other trends appear in Figure 9. RosettaDOM (just 4 boundaries predicted in the Remote regime, 2 correct at ±10 residues) tends to under-predict boundaries, and therefore has reasonable specificity but poor sensitivity. Ginzu_SEQ (3 predictions, 1 correct at ±10 residues), applied to the low-confidence 3D-Jury targets (4 targets with 8 official internal boundaries), does not fare well, but DP_Hybrid (9 predictions, 5 correct at ±10 residues, 4 of which are still correct at ±1 residue) exhibits a good balance between the conservative predictions of Ginzu_SEQ and RosettaDOM and the aggressive predictions of Ginzu_HM.

Comparison of the different approaches of Ginzu_HM, Ginzu_TP, Ginzu_SEQ, RosettaDOM, and DP_Hybrid supports the common sense approach of using the best available method for a given difficulty regime. Consensus structure-based parsing of the parent structure and homologs is sufficient in the PSI-BLAST regime and when a confident PDB homolog is detected by fold recognition, consensus structure-based parsing of the best possible model is superior. Hybrid approaches that find agreement between low-confidence fold recognition and *de novo* models and sequence-derived signals tend to yield the most stable solutions for the most remote targets.

CASP6/CASP7 comparisons

CASP6 was the first CASP experiment in which domain prediction was assessed as a separate category; so it is important to measure progress in domain prediction. It is clear that the mean NDO scores for the targets in CASP7 are higher than those of CASP6, but the comparison between the two experiments is not that straightforward because the difficulty of predicting the domain boundaries for each target needs to be taken into consideration and the CASP7 targets may have been easier to predict.

First CASP6 and CASP7 targets have to be ranked in terms of prediction difficulty. Prediction difficulty depends on a number of factors. One of the most important factors is whether a template can be found to model the structure, but other factors such as the length of the protein, the similarity to known structures, the compact-

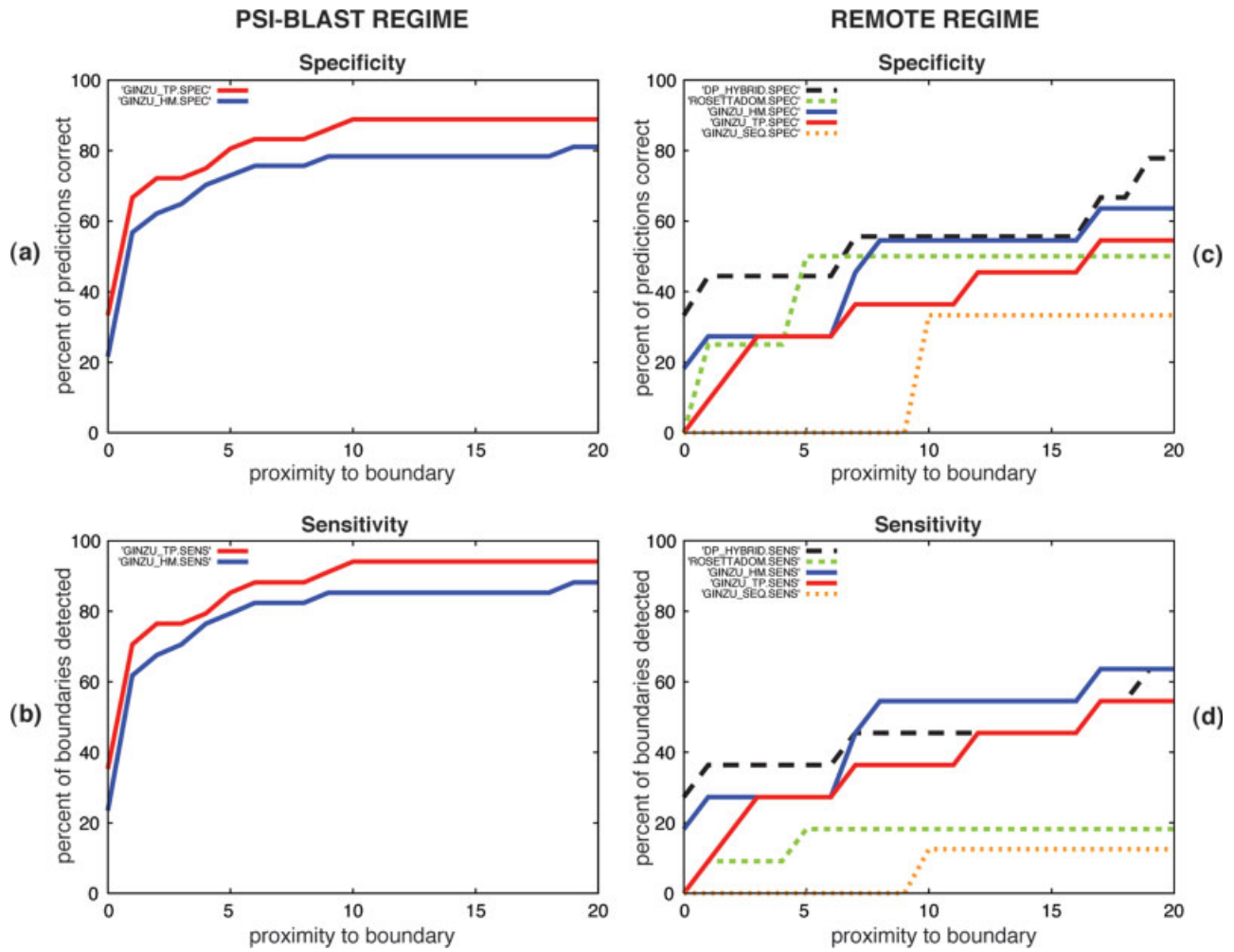

Figure 9

Boundary prediction for multiple domain targets by Ginzu, RosettaDOM, and DP_Hybrid. The "specificity" and "sensitivity" of boundary predictions for multiple domain targets by Ginzu_HM (blue line), Ginzu_TP (red line), Ginzu_SEQ (dashed yellow line), RosettaDOM (dashed green line), and DP_Hybrid (dashed black line). Correct boundaries are defined as being within a sequence deviation tolerance (x-axis) of the assessor-defined boundary. Targets are separated into difficulty regimes, with easy targets ("PSI-BLAST regime") defined as those with a confident PDB homolog detected by Robetta using PDB-BLAST and harder targets ("Remote" regime) as those without a PSI-BLAST detectable PDB homolog. The Remote regime comprised just targets T0299, T0301, T0321, T0347, T0356, and T0372. In the PSI-BLAST regime Ginzu_SEQ was not used, while RosettaDOM and DP_Hybrid used Ginzu_HM predictions. Ginzu_SEQ did not predict targets T0347, T0301 and T0321 in the Remote regime. The domain definitions used for the figures were the first alternative domains defined by the domain category assessors. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

ness of the domains in the structure and whether the domains are continuous or non-continuous. The number of domains in each structure is also a factor and of course single domains are easiest to predict—the mean NDO score of the best 10 predictors for those targets defined as single domains in CASP6 was 96.55, while the mean for the best 10 predictors for single domain targets in CASP7 was an incredible 99.98.

We also generated domain predictions from the target structures using the protein structure domain-parsing program PDP. The NDO scores were calculated for each of the domains parsed with PDP as if the program were just another predictor. We used the PDP NDO scores in two ways, first as a means of assessing the difficulty of predicting domains for a target structure and second as a control to allow us to compare prediction groups in CASP6 and CASP7.

Target difficulty ranking were calculated for the 17 CASP6 and 31 CASP7 multiple domain targets. Targets were ranked by the percentage identity to the nearest structural template, by their PDP NDO scores and by a crude domain difficulty ranking that took into account the number of domains in the protein and whether the domains were continuous or non-continuous (non-continuous domains in general are harder to predict). The final target difficulty ranking was simply the mean of the

 DOI 10.1002/prot

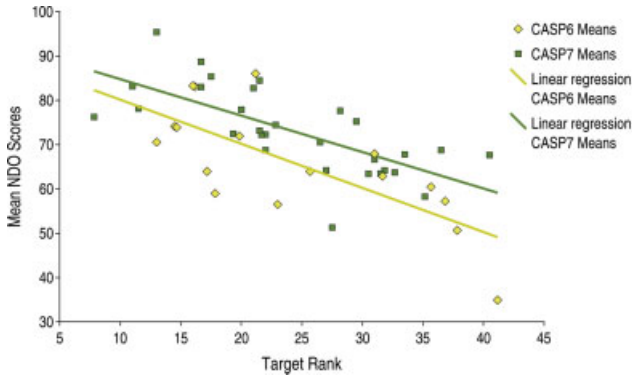

Figure 10

CASP6 and CASP7 compared. Mean NDO scores for each target from CASP6 and CASP7. NDO scores are the mean of the best 10 predictors for each target. Scores are plotted against target rank as explained in the text and regression lines have been plotted on the figure for the sets of scores from each experiment. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

three sets of rankings. With the CASP6 and CASP7 targets ranked in this way we were able to plot the NDO scores for each target against target difficulty ranking (Fig. 10). The plot shows quite convincingly that the higher scores obtained in CASP7 were not just a consequence of the targets being easier to predict.

We also used the PDP parsed domain as a control to compare the groups in each experiment. PDP was regarded as just another predictor and the NDO scores for the PDP parsed domains were compared to the NDO scores of each of the predictors in both competitions. Groups and the PDP parsed domains were compared head to head using paired t-tests. The results for the two experiments can be seen in Figure 11.

In CASP7 there were six groups (DP722, DP581, DP497, DP136, DP556, and DP105) that could not be distinguished statistically from PDP, one of the best structural parsing programs, while in CASP6 just two groups had *P* values that were not statistically significant. However, the *P* values for the two best CASP6 groups were very low, almost borderline for significance. These two CASP6 groups (DP353 and DP421) were groups DP581 and DP497 in CASP7. The fact that there are 6 groups with scores statistically similar to PDP in CASP7 (even though there were almost twice as many multiple domain targets in CASP7) is further confirmation of the general improvement in domain prediction between the two CASP experiments.

CONCLUSIONS

Domain prediction was added as a new assessment category in CASP6 and it is clear that there has been a general improvement in prediction since the first experiment. However, it is not clear how much of this improvement is due to real advances in methods and how much stems from groups converging on a set of reliable methods. It is also possible that some of the improvement may be down to predictors adapting to the scoring scheme used for CASP6 and again in CASP7.

This assessment has shown that a number of groups seem to be able to make reliable, good quality template-based domain predictions. Where it is possible to build a model of the target structure, a robust domain-parsing program such as PDP will usually be able to make a good approximation of the domain boundaries. And as usual the better the alignment and model, the more accurate the domain boundary predictions tend to be.

Although the standard of domain boundary prediction was generally quite high, even the best-scoring prediction methods struggled to predict the exact domain boundaries for some of the harder multidomain targets, particularly those targets with noncontinuous domains and those targets where domain boundaries fell in regions that had to be modeled *ab initio* (for example targets T0299, T0289).

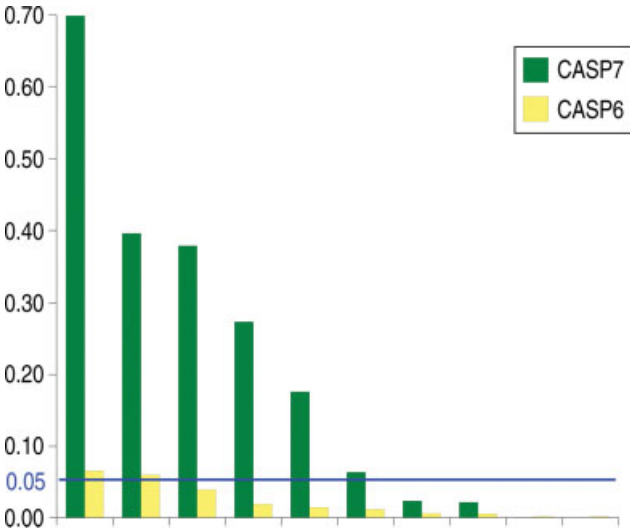

Figure 11

Paired t-tests between PDP and predicting groups from CASP6 and CASP7. The chart shows the P *values from the paired* t*-tests for NDO scores between each of the CASP6 and CASP7 participating groups and the "PDP predictor". The PDP predictor was simply the domains predicted from the structure-based domain predictor, PDP. The* P *value 0.05 is the cut-off for significance for the comparisons and is marked by a horizontal line. Groups with p-values higher than the line are not significantly different from the PDP predictor. In CASP6 two groups have* P *values that are very close to the limit for significance, but six predicting groups in CASP7 have* P *values that suggest that there is no statistical difference between their predictions and those of the structure-based domain parser. It should be noted that there are two CASP6 groups with higher* P *values that do not appear in the chart, but these groups made predictions for just 2 and 6 multiple domain targets. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]*

The inclusion of domain prediction in the CASP experiment has shown that the best template-based domain prediction methods are of a high standard. However, it has also become obvious that CASP is not the best format for the evaluation of domain prediction. There were few free modeling targets in the experiments and no difficult multidomain free modeling targets at all. We were not able to evaluate the standard of *ab initio* domain prediction in this CASP.

The targets submitted to CASP for structure prediction are not sufficiently challenging for ab initio domain prediction and there are unlikely to be large numbers of difficult to predict targets in future CASP experiments. Many of the target proteins came from structural genomics projects and it seems that structural genomics projects have a tendency to select single domain proteins. The CASP domain prediction category might have a future if it concentrates wholly on template-based predictions, but at the same time the best template-based methods have shown that they have already reached a high standard and can compete with the best structural parsing methods.

Where there is room for improvement appears to be in *ab initio* prediction, although even this statement is difficult to justify with evidence because there have been so few difficult multidomain targets in the last two CASP experiments. The infrequency with which difficult multidomain targets are deposited in the PDB means that it would make much more sense to carry out any future blind testing of *ab initio* domain prediction servers in a continuous format similar to the rolling assessments that were implemented in the EVA[38] and LiveBench[39] experiments.

ACKNOWLEDGMENTS

We thank all who helped at the Prediction Center, especially Andriy Kryshtafovych and the CASP organizers. We thank fellow assessors Randy Read, Neil Clarke, and Torsten Schwede for their input on the domain definitions and Alfonso Valencia for invaluable assistance. We are indebted to Chin-Hsien Tai and BK Lee for use of the CASP6 prediction data.

REFERENCES

1. Shiozawa K, Maita N, Tomii K, Seto A, Goda N, Akiyama Y, Shimizu T, Shirakawa M, Hiroaki H. Structure of the N-terminal domain of PEX1 AAA-ATPase. Characterization of a putative adaptor-binding domain. J Biol Chem 2004;279:50060–50068.
2. Arakaki T, Le Trong I, Phizicky E, Quartley E, DeTitta G, Luft J, Lauricella A, Anderson L, Kalyuzhniy O, Worthey E, Myler PJ, Kim D, Baker D, Hol WG, Merritt EA. Structure of Lmaj006129AAA, a hypothetical protein from Leishmania major. Acta Crystallogr Sect F Struct Biol Cryst Commun 2006;62:175–179.
3. Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P. SMART 5: domains in the context of genomes and networks. Nucleic Acids Res 2006;34:D257–D260.
4. Liu J, Rost B. Domains, motifs, and clusters in the protein universe. Curr Opin Chem Biol 2003;7:5–11.
5. Levitt M, Chothia C. Structural patterns in globular proteins. Nature 1976;261:552–558.
6. Tai C-H, Lee W-J, Vincent JJ, Lee BK. Evaluation of domain prediction in CASP6. Proteins 2005;61:183–192.
7. Clarke ND, Ezkurdia I, Kopp J, Read RJ, Schwede T, Tress ML. Domain definition and target classification for CASP7. Proteins 2007;69(Suppl 8):10–18.
8. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
9. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
10. Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grant A, Lee D, Akpor A, Maibaum M, Harrison A, Dallman T, Reeves G, Diboun I, Addou S, Lise S, Johnston C, Sillero A, Thornton J, Orengo C. The CATH domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res 2005;33:D247–D251.
11. Kim DE, Chivian D, Malmstrom L, Baker D. Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. Proteins 2005;61:193–200.
12. Cheng J, Baldi P. A machine learning information retrieval approach to protein fold recognition. Bioinformatics 2006;22:1456–1463.
13. Vapnik V. Statistical learning theory. New York: Wiley; 1998.
14. 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.
15. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
16. Edgar R, Sjolander ,K. COACH: profile–profile alignment of protein families using hidden markov models. Bioinformatics 2004;20:1309–1318.
17. Marti-Renom MA, Stuart A, Fiser A, Sánchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325.
18. Alexandrov N, Shindyalov IN. PDP: protein domain parser. Bioinformatics 2003;19:429–430.
19. Cheng J, Sweredoski MJ, Baldi P. DOMpro: protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks. Data Mining Knowl Discov 2006;13:1–10.
20. Baldi P, Pollastri G. The principled design of large-scale recursive neural network architectures-DAG-RNNs and the protein structure prediction problem. J Mach Learn Res 2003;4:575–602.
21. Kraulis PJ. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr 1991;24:946–950.
22. Holland T, Veretnik S, Shindyalov IN, Bourne PE. A benchmark for domain assignment from protein 3-dimensional structure and its applications. J Mol Biol 2006;361:562–590.
23. Marsden RL, McGuffin LJ, Jones DT. Rapid protein domain assignment from amino acid sequence using predicted secondary structure. Protein Sci 2002;11:2814–2824.
24. Liu J, Rost B. Sequence-based prediction of protein domains. Nucleic Acids Res 2004;32:3522–3530.
25. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.
26. Wang G, Dunbrack RL, Jr. PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res 2005;33:W94–W98.
27. Lee J, Scheraga HA, Rackovsky S. New optimization method for conformational energy calculations on polypeptides: conformational space annealing. J Comput Chem 1997;18:1222–1232.

 DOI 10.1002/prot

28. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714–2726.
29. Sim J, Kim SY, Lee J. PPRODO: prediction of protein domain boundaries using neural networks. Proteins 2005;59:627–632.
30. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.
31. Chivian D, Kim DE, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss CE, Bonneau R, Rohl CA, Baker D. Automated prediction of CASP-5 structures using the Robetta server. Proteins 2005;53:524–533.
32. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulate anealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
33. Taylor WR. Protein structural domain identification. Protein Eng 1999;12:203–216.
34. George RA, Heringa J. SnapDRAGON: a method to delineate protein structural domains from sequence data. J Mol Biol 2003;316:839–851.
35. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acids Res 2004;32:D138–D141.
36. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
37. Chivian D, Baker D. Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. Nucleic Acids Res 2006;34:e112.
38. Koh IYY, Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Eswar N, Graña O, Pazos F, Valencia A, Sali A, Rost B. EVA: evaluation of protein structure prediction servers. Nucleic Acids Res 2003;31:3311–3315.
39. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. LiveBench-1: continuous benchmarking of protein structure prediction servers. Protein Sci 2001;10:352–361.