Evaluation of domain prediction in CASP6

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · JANUARY 2005

Impact Factor: 2.63 · DOI: 10.1002/prot.20736 · Source: PubMed

Evaluation of Domain Prediction in CASP6

Chin-Hsien Tai,[1] Woei-Jyh Lee,[2] James J. Vincent,[1] and Byungkook Lee[1]*
[1]*Laboratory of Molecular Biology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland*
[2]*Department of Computer Science, University of Maryland, College Park, Maryland*

ABSTRACT **We present an analysis of the domain boundary prediction, a new category, in the sixth community-wide experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP6). There were 1011 predictions submitted for 63 targets. Each prediction was compared to the set of domains defined manually by visual inspection of the experimental structure. The comparison was scored using a new domain prediction scoring scheme. As the definition of a domain is subjective, many targets were assigned alternate definitions. For such targets, each prediction was compared with all different definitions and the best score was chosen. The predictors found it difficult to accurately predict domain boundaries when the target protein contained many domains or domains made of multiple sequence segments. The CBRC-DR (P0536) and Sternberg (P0237) groups were the most successful among human experts, while Baker-Rossettadom (P0353) and Baker-Robetta-Ginzu (P0421) did well among servers. Proteins 2005;Suppl 7:183–192.**

Key words: domain; prediction; assessment; protein structure; CASP6

INTRODUCTION

Domains are basic units of protein structure, and each is often associated with a specific function. Protein structures are usually classified in units of domains.[1,2] Identification of the domain boundaries in the absence of structural information is an important problem in structural biology. Domain boundaries can be used to cut proteins in biochemical experiments for functional and structural studies. Domain boundary prediction is usually the first step in protein structure prediction.

Many methods for predicting domain boundaries rely on the result of protein multiple sequence alignment.[3–6] However, domain size,[7] side chain entropy,[8] predicted secondary structures,[9] the difference in the frequency of amino acids observed in the domain and linker regions (linker index),[10] the neural networks,[11–13] and the hidden Markov models[14,15] have also been used to predict domain boundaries. As in the case of structure prediction, it is important that the prediction results be evaluated objectively to identify effective techniques to facilitate further progress. Recognizing the importance of this problem, the CASP6 organizers added domain prediction as a new category of objective assessment in 2004. This is the report of the assessment of these predictions. CAFASP (Critical Assessment of Fully Automated Structure Prediction), a parallel evaluation service for automatic servers, has also added domain prediction as a new category in CAFASP4 (http://www.cs.bgu.ac.il/~dfischer/CAFASP4/index.html).

The first step for the assessment was to define the "true" domains. This was done manually by visual inspection of each target protein structure (by B.L.). In the majority of cases, this task was straightforward and the protein could be parsed into a unique set of domains. In many other cases, however, it was felt that the protein structure could be broken into domains in more than one way. In such cases, the protein was assigned a set of multiple domain definitions, one of which was the "official" definition given at the CASP6 Web site.[16] A prediction was compared to all definitions, both "official" and alternates, and the best score was used for evaluation.

Several different scoring schemes have been used in the past. Two different measures, the number of domains predicted and an overlap score between the predicted and "real" domains, were used in CAFASP4 (http://www.cs.bgu.ac.il/~dfischer/CAFASP4/index.html). A different, common method is to make a binary decision to consider a predicted domain boundary to be correct if it is within ±20 residues from a defined boundary.[5,7,9,12,13] Under this criterion, however, two predictions can both be considered correct even when the predicted sizes of a domain varies by up to 80 residues, if the domain is bounded by two boundaries on either side. Gracy and Argos[3] used overlap and coverage measures, which are the number of residues that are common in both defined and predicted domains divided by the number of residues in the union of the defined and predicted domains (overlap) or in the defined domain only (coverage). These are good measures when the predicted domain can be unambiguously associated with a defined domain. For the CASP predictions, however, it is often difficult to know which predicted domain should be considered to be a prediction of which defined domain. When the association between predicted and defined domains is uncertain and must be determined

*Correspondence to: Byungkook Lee, Bldg. 37, Room 5120, 37 Convent Dr., MSC 4264, Bethesda, MD 20892-4264. E-mail bk@nih.gov

Received 20 May 2005; Accepted 24 June 2005

Published online 26 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20736

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

TABLE I. Overlap Matrix for the Prediction P0237 for Target T0199

Def. \ Pred.	D1[a] (14–98)	Linker1 (99–103)	D2a (104–142)	Linker2 (143–144)	D3 (145–226)	Linker3 (227–229)	D2b[a] (230–336)
d1 (1–98)	85	0	0	0	0	0	0
Linker (99–115)	0	5	12	0	0	0	0
d2 (116–338)	0	0	27	2	82	3	107

[a]Residue 1–13 and 337–338 are missing.

first, it becomes complex at best to use two, rather than a single, quality measures. Nagarajan and Yona[11] associated each predicted boundary with the closest defined boundary and used four different measures, two of which measure the distance between predicted and defined boundaries and the other two based on binary decision using 10-residue cutoff. Aside from the problem of handling so many measures, associating each predicted boundary to the closest defined boundary can result in inconsistencies and unfair associations for some predictions. Marsden et al.[9] and Liu and Rost[12] used the number of domains and the binary decision using 20-residue cutoff.

Use of the number of domains as a quality measure presents at least two problems for evaluating the CASP predictions. One problem is that the true number of domains is not known for some target structures because they have missing residues, which may or may not form a separate domain. The other problem is that, for many target proteins, multiple definitions of domain structure are possible so that two predictions with different number of domains can both be considered to be correct.[16]

We experimented with several schemes but in the end settled for a new method that does not use an ad hoc parameter. It is based upon overlap score and generates a single score as the measure of the quality of a prediction, which includes explicit penalties for both under- and overpredictions (too small and too large, respectively) of domains. The association between predicted and defined domains is made on the basis of maximal overlap. The penalty for predicting a wrong number of domains is included implicitly by the penalty associated with an extra or missing domain boundary.

Average scores were analyzed in several different ways to gain insight into which groups performed better and what brought the difficulties. Finally, manual inspection was also performed to verify the performance of the prediction group and of the new scoring scheme.

METHODS

Domain Prediction Scoring Scheme

A mathematical description of the scoring scheme is available as a supplementary material on the Web site. Here we describe the procedure using a prediction from group P0237 for an alternate domain definition of the target T0199. Target T0199 is 338 residues long and has three domains. The second domain (D2) is made of two segments (D2a and D2b), which are separated in sequence by two linkers and the third domain (D3). The N-terminal 13 residues and the C-terminal two residues are missing in the target structure. The prediction is made of two domains, d1 and d2, separated by a 17-residue linker. The boundaries of the defined and predicted domains are given in the first row and column of Table I.

TABLE II. Consolidated Overlap Matrix and Net Scores for the Table I Example

	Linkers	D1	D2	D3	Score
linker	5	0	12	0	
d1	0	85	0	0	85
d2	5	0	134	82	47
score		85	122	82	210.5

Step 1

Compute the overlap between each segment—domain segments, leader, linkers, and trailer—of the target definition and the prediction. Missing residues are not included. This generates a 3×7 matrix shown in Table I.

Step 2

Consolidate the matrix by summing up overlaps for all segments of each domain into one column or row and for all linkers, leader, and trailer into one column or row. The size of the consolidated matrix (Table II) is one (for the linkers) plus the number of domains in each dimension.

Step 3

The net overlap score is computed from the consolidated matrix for each of the target domains and for each of the predicted domains. To compute the score for the target domain D2, for example, the largest element in the D2 column is selected and all other numbers in the same column are subtracted from it. This number, $134 - 12 = 122$, is written in the bottom row of Table II. Similarly, the score for the predicted domain d2 is 134, the maximum element in the d2 row, minus all other entries of the row, $5 + 82$. This result is written in the last column of Table II. No score is computed for the linkers in either the target or the prediction. Thus, a correct prediction of the linkers (the matrix element 1,1) is not explicitly recognized. The total raw score of the prediction is the sum of the scores for each predicted and defined domains divided by 2. This is shown at the bottom right-hand corner of Table II.

Step 4

A normalized score is obtained by dividing the raw score by the perfect score. The latter is the total number of residues in the defined domains, which in this case is 313. The final score is, therefore, $210.5/313 = 67.25\%$. In this case, this is essentially the fraction of residues in d1–D1

and d2–D2 overlaps. We call this the Normalized (or Net) Domain Overlap (NDO) score.

Overall Score for Each Group

To quantify the overall performance of each predicting group, four different measures were calculated. These are the average NDO score, average *Z*-score, the number of best predictions (top-1 frequency) and the number of prediction scores within 5% of the top score (top-score frequency). Averages were computed across different targets only over the submissions. Nonsubmissions were ignored rather than given an arbitrary score.

We also looked at the three-dimensional structures of each prediction using PyMol[17] to verify the calculations in individual cases.

RESULTS

Domain Definitions

There were 76 targets for which the coordinates became available in time for the prediction assessment, but 13 of these were canceled for various reasons. The defined domain boundaries for the 63 targets are given in Supplementary Table I on the Web site. There were 27 targets for each, of which more than one set of domains were defined.

We use the capital letter D followed by a serial number to designate the domains in a definition. The serial number usually indicates the order of the domain in the sequence. Many domains contain segments that are not contiguous in the sequence. These are indicated by appending a segment-specific lower case letter to the name. The domains in a prediction are designated similarly, except that the lower case letter d is used as the first character of the name.

Prediction Scores

We evaluated 1011 predictions from 22 groups for the 63 targets. Eleven groups were servers and the rest were human experts. The number of predictions each group submitted varied from only one to all 63 (see Table IV). On the other hand, the number of predictions for each target was similar; 87% of them had 15 to 19 predictions (see Table III).

The Normalized Domain Overlap (NDO) score was calculated for each prediction against all alternate definitions of the target and the best score was taken. The NDO scores for all the predictions are in Supplementary Table II. Unless otherwise noted, a score means the NDO score throughout this article.

Easy and Difficult Targets

Table III lists the NDO scores averaged over all predictions for each target, as well as the best scores for each target. The targets were sorted in ascending order of the number of domains. There were 29 targets, for which at least one prediction had the full score. All but two of these had one-domain definition, on which the full score was invariably based. For all but four targets, at least one group had a score greater than 90%. The average scores for one-domain targets were above 80% except for two targets. The average score then decreases as the number of domains in the target increases. Difficult targets with the average NDO score below 60% are marked in Table III. Prediction results are described below for a few targets.

T0248

The target that had the lowest average and lowest best score was T0248. This structure is made of three domains (Fig. 1). Except for one β-hairpin in D2, all three domains are basically helix bundles of similar length. From the pattern of the secondary structures alone, it would be difficult to know if a loop between two helices is intra- or interdomain for such a structure. Eleven of the 17 predictions were one-domain, for which the NDO score was 27.9%. Three other predictions were two-domains and another a four-domain. There were two three-domain predictions. One had the second worst score of 35.4%, while the other had the highest score of 61.1%. Both had an expanded d2 with the two domain boundaries inside of D1 and D3, but the former had them more deeply inside both flanking domains. The better prediction from the Oka group (P0461) had the d1–d2 boundary nearly correct, but the d2–d3 boundary was placed one and a half helices inside D3 [Fig.1(B)].

T0237

T0237 is an officially three-domain protein, but the three domains are poorly separated geometrically. It has a long N-terminal leader that spans all three domains, many missing residues, and a C-terminal trailer that interacts with D2 [Fig. 2(A)]. The top scoring prediction (P0381) had the d1–d2 and d2–d3 boundaries at nearly correct positions, but included the N-terminal leader in d1 and the C-terminal trailer in d3 [Fig. 2(B)]. An alternate definition for this target combines D2 and D3 of the official definition [Fig. 2(C)]. The second high scoring prediction (P0089) was a two-domain [Fig. 2(D)]. It had a d1–d2 boundary near the D1–D2 boundary, but the N-terminal leader was in d1 and the C-terminal trailer was in d2, which was made of both D2 and D3 as in the alternate definition. The third high scoring prediction was one-domain. The closely following fourth was a two-domain with basically D1 and D2 combined.

T0268 and T0279

Each of these has two domains, one of which is made of N- and C-terminal two segments. Many predictors did poorly because they did not recognize the two segments of D1. On the other hand, both proteins are made of CM fold class domains, which means that similar structures can be found in the database by a sequence homology search. This is probably why there are nearly perfect top-scoring predictions (P0353 and P0421).

T0241

Target T0241 is another example of a protein with domains made of multiple segments. This is officially a two-domain protein in which each domain is made of three segments in D1a–D2a–D1b–D2b–D1c–D2c sequence. Both

TABLE III. Summary of Prediction Score for Each Target

Target	Nd[a]	Fold class[b]	Np[c]	Mean score	Best score	Nt[d]
T0196	1	CM/hd	13	96.16	98.19	1
T0197	1	FR/H	13	82.68	96.45	6
T0201	1	NF	16	90.42	97.78	11
T0203	1	FR/H	12	79.55	97.41	6
T0205	1	CM/hd	11	93.85	99.49	1
T0206	1	FR/H	15	76.79	98.11	1
T0208	1	CM/hd	12	88.09	100.00	5
T0211	1	CM/hd	16	94.69	100.00	11
T0212	1	FR/A	16	87.42	100.00	1
T0215	1	FR/A	16	93.68	94.79	15
T0224	1	FR/H	16	95.46	98.21	14
T0227	1	FR/H	16	89.70	100.00	11
T0230	1	FR/A	17	82.81	93.33	11
T0231	1	CM/ez	16	92.04	100.00	10
T0234	1	CM/hd	17	90.25	98.47	13
T0238	1	NF	16	89.52	97.38	12
T0240	1	CM/ez	17	86.40	93.75	11
T0242	1	NF	17	88.52	100.00	12
T0243	1	FR/H	15	95.11	100.00	11
T0251	1	FR/H	17	97.00	100.00	13
T0263	1	FR/H	17	86.95	94.83	1
T0265	1	CM/hd	17	90.02	100.00	10
T0271	1	CM/ez	17	83.51	90.37	12
T0274	1	CM/ez	16	88.39	97.30	11
T0275	1	CM/ez	16	96.67	100.00	14
T0276	1	CM/ez	15	84.76	100.00	8
T0277	1	CM/ez	15	96.30	100.00	14
T0281	1	FR/A	15	93.71	100.00	12
T0282	1	CM/ez	15	86.81	100.00	6
T0198	1 or 2	FR/A	17	86.07	99.55	1
T0200	1 or 2	CM/hd	15	82.20	100.00	10
T0204	1 or 2	CM/ez	13	86.17	100.00	1
T0213	1 or 2	FR/H	16	89.96	100.00	1
T0214	1 or 2	FR/H	16	92.09	100.00	1
T0226	1 or 2	CM/hd, CM/hd	18	78.25	99.64	1
T0229	1 or 2	CM/ez, CM/ez	16	88.94	100.00	10
T0232	1 or 2	CM/hd, CM/hd	18	85.56	100.00	5
T0239	1 or 2	FR/A	15	91.80	100.00	10
T0241	1 or 2	NF, NF	18	64.70	100.00	2
T0244	1 or 2	CM/ez	16	81.62	96.57	8
T0246	1 or 2	CM/ez	17	85.15	100.00	11
T0266	1 or 2	CM/ez	17	89.39	100.00	13
T0267	1 or 2	CM/hd	16	84.32	100.00	8
T0273	1 or 2	NF	16	84.33	100.00	9
T0280	1 or 2	CM/ez, FR/A	15	79.58	100.00	5
T0235	1 or 2 or 3	CM/ez, FR/A	19	74.70	99.59	1
T0209	2	FR/A, NF	13	86.00	96.95	1
T0222	2	CM/hd, FR/H	18	83.06	97.10	1
T0223	2	CM/hd, FR/H	17	62.87	97.98	1
T0228^*	2	FR/H, FR/H	17	57.25	88.98	2
T0233	2	CM/ez, CM/ez	19	83.26	100.00	1
T0249	2	FR/H, FR/H	19	63.95	96.00	2
T0262	2	FR/A, FR/H	17	71.91	96.45	1
T0264	2	CM/ez, CM/hd	18	73.88	100.00	4
T0268^	2	CM/ez, CM/ez	14	58.99	98.58	2
T0269	2	CM/ez, CM/hd	16	70.55	91.10	1
T0272	2	FR/A, FR/A	17	63.94	98.37	1
T0279^	2	CM/hd, CM/hd	16	56.51	97.98	2
T0237^*	2 or 3	FR/H, FR/H, FR/H	18	50.66	79.08	1
T0247	2 or 3	CM/ez CM/ez CM/ez	16	74.09	98.61	2
T0216	2 or 3 or 4	NF, NF	18	67.93	98.34	1
T0199^*	3	CM/hd, FR/H, FR/A	16	60.46	73.11	1
T0248^*	3	FR/A, NF, FR/A	17	34.94	61.07	1

[a]Number of domains.
[b]Fold class categories refer to officially defined domains. CM/ez: "easy" comparative modeling target; CM/hd: "hard" comparative modeling target; FR/H: homologous fold recognition target; FR/A: analogous fold recognition target; NF: new fold target.
[c]Number of predictions submitted for each target.
[d]Number of ties for the best NDO score.
^These have average NDO score at or below 60%.
*These have the best NDO score below 90%.

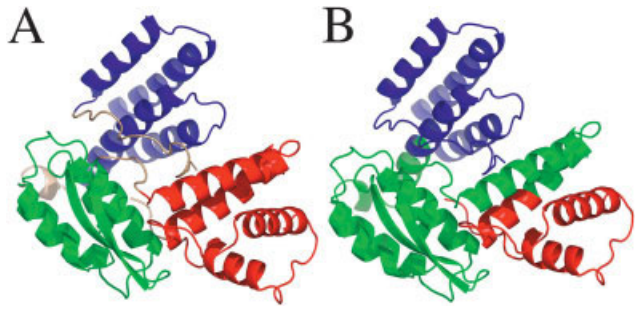

Fig. 1. Target T0248 (**A**) and the prediction (**B**) from P0461. The three domains D1, D2, and D3 are colored blue, green, and red, respectively, in (**A**) and linkers are in wheat color. The corresponding predicted domains are similarly colored (**B**).

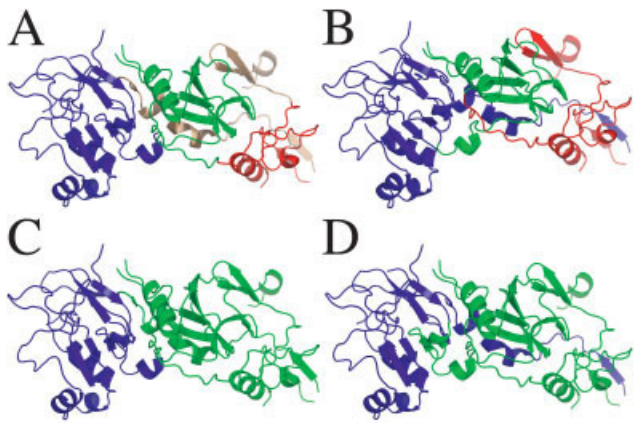

Fig. 2. Target T0237. Domains D1, D2, and D3 are colored in blue, green, or red, and the leader, trailer, and linkers are in wheat color. (**A**) Official three-domain definition. (**B**) The top scoring prediction (P0381) includes the leader in the first domain (blue) and the trailer in the third domain (red). (**C**) Alternate two-domain definition of T0237. (**D**) The second best prediction (P089).

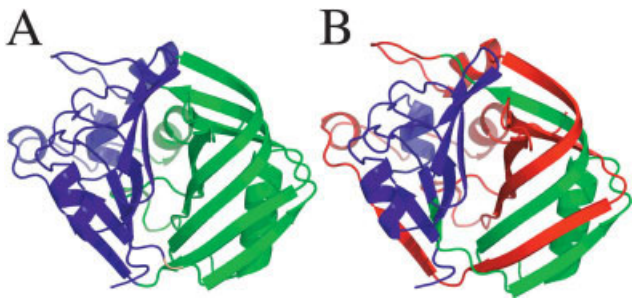

Fig. 3. Target T0241 (**A**) and the prediction (**B**) from P0353. Different colors indicate different defined or predicted domains.

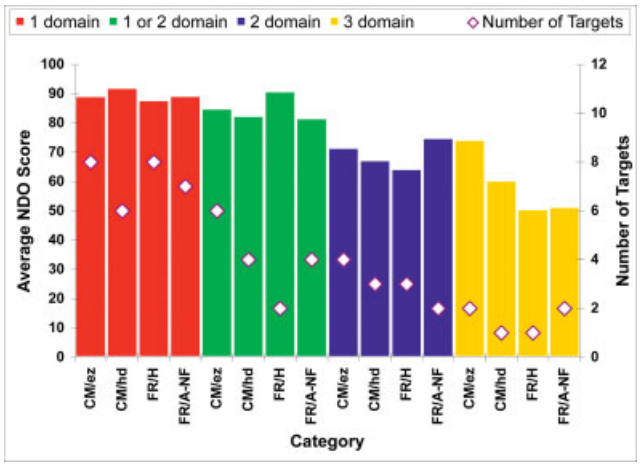

Fig. 4. The NDO scores averaged for different target categories. There are four fold class categories: CM/ez, CM/hd, FR/H, and FR/A-NF combination. There are also four target domain number categories: one-domain, one- or two-domain, twp-domain, and the rest, which is labeled as three-domain. The combination yields $4 \times 4 = 16$ target categories. The NDO values averaged over each of these different categories are plotted as a bar graph. For easy viewing, bars are colored according to the number of domain categories, as indicated in the legend at the top of the graph. The blank diamonds give the number of targets in each category using the right-hand side scale.

domains are among the most difficult NF targets[18] and, again, no prediction had either of these domains correctly predicted. The average NDO score is below 65 (Table III) even though the alternate one-domain definition rewarded one-domain predictions with the full score. Figure 3 shows the target structure in panel A and the prediction from the group P0353 with score 44.7% in panel B as an example. Both domains in panel B are in a mosaic of colors, indicating a serious error in domain prediction. Yet the first two of the five domain boundaries in this structure were predicted excellently well. The main problem was in missing, not misplacing, a couple of boundaries, between D1b–D2b and D2b–D1c.

Domain Number, Fold Class, and Length Dependence

It may be expected that domain structure prediction is easier for the targets for which the 3D structure can be predicted for some of its domains. To test this possibility, the targets were tagged according to the fold class designation of their official domains. There are five fold class designations: CM/easy, CM/hard, FR/H, FR/A, and NF.[16] We pooled FR/A and NF domains and assigned each protein one of the four class labels according to the class designation of the easiest domain that the protein contains. Figure 4 shows the NDO scores, averaged and sorted according to these fold class labels and by the number of domains of the targets.

Both Figure 4 and Table III show that the average NDO score decreases as the number of domains increases. In terms of the fold class dependence, the average NDO score is independent of the fold class for one-domain targets but, for others, it generally decreases as one moves from the CM/easy to FR/A-NF fold classes. However, the small sample size makes it difficult to be certain if the trend is real. The visual inspection of targets T0268 and T0279 suggests that good predictors did take advantage of the ease with which a similar structure could be found in the database, but many others did not.

We also examined the dependence of the average NDO scores on the size (number of amino acids) of the target (Fig. 5). There is very little length dependence when the targets are less than about 200 residues long, in which

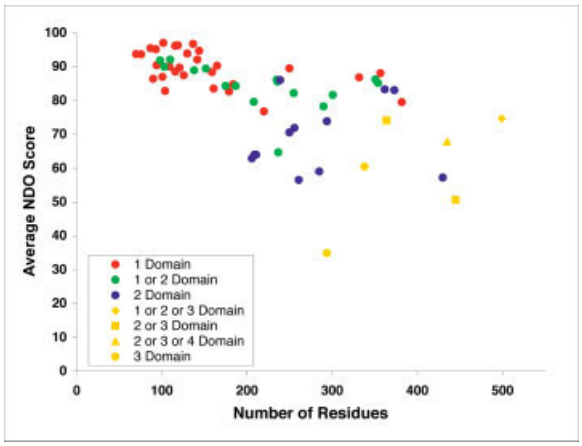

Fig. 5. The average NDO scores of each target plotted against the amino acid length of the target protein. Each point is color coded as in Figure 4.

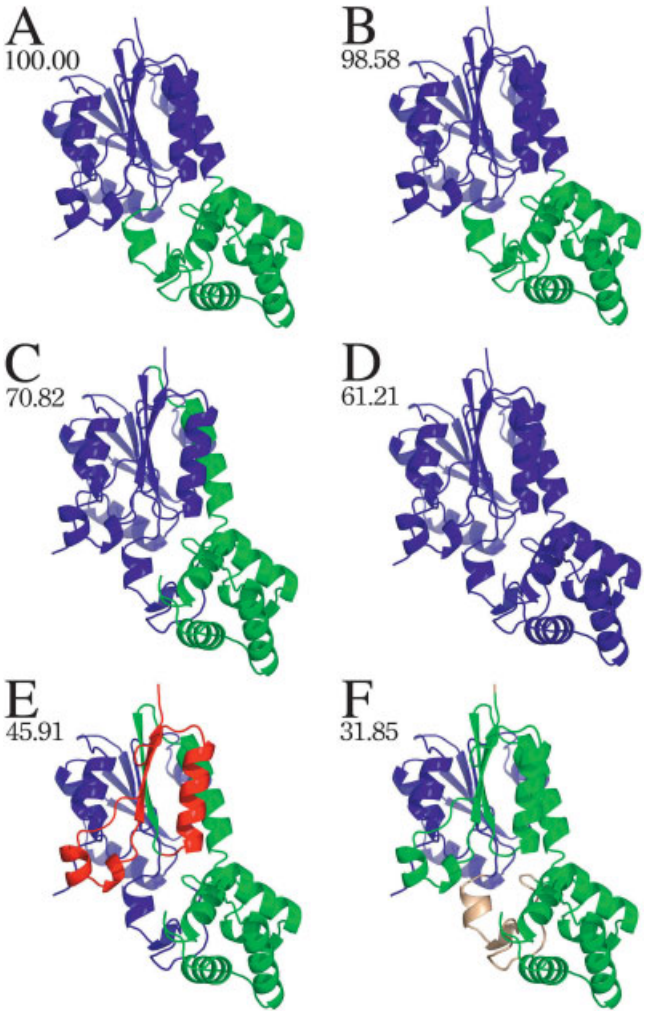

Fig. 6. The relationship between NDO scores and the prediction quality for the two-domain target T0268. The scores are indicated in each panel and the domains are colored in blue, green, or red. (**A**) The domain definition of target T0268. The blue domain is made of two segments. (**B**) Prediction by P0353. (**C**) Prediction by P536. (**D**) One-domain prediction by P089. (**E**). Three-domain prediction by P682. (**F**) Prediction by P667. The long linker between the two domains is in wheat color.

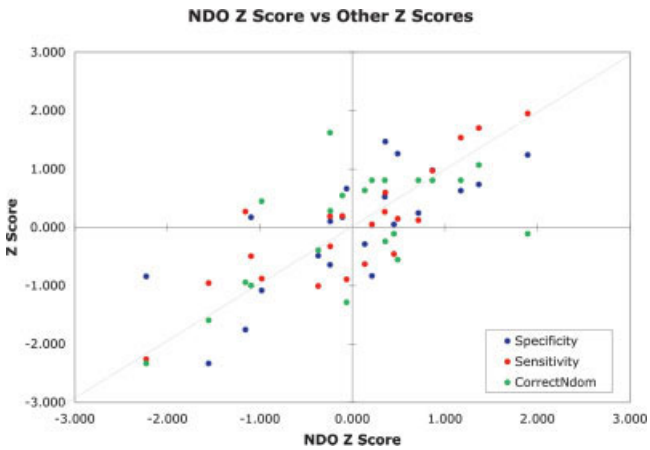

Fig. 7. The *Z*-score of the specificity (blue dots) and sensitivity (red dots) of domain boundary predictions and the *Z*-score of the average number of targets for which the number of domains were correctly predicted (green dots) plotted against the *Z*-score of the NDO scores for all predicting groups. The average is taken over the predictions made by each group among the 17 targets, which consists of more than one domain. For targets with multiple definitions, the definition that gave the best score was used for each scoring scheme. If the three scoring schemes were perfectly correlated with the NDO score, all points would fall on the diagonal line indicated.

case they are either one-domain or one- or two-domains. Beyond 200 residues, there is a downward trend with length, but the number of domains also generally increases with length. Among the targets with the same number of domains, the length dependence is either absent or positive. Again, the small sample size makes it difficult to be certain, but length per se does not appear to be a strong factor in determining the difficulty of a target.

Prediction Group Rankings

To quantify the overall performance of each predicting group, four different measures were calculated as described in Methods. The results calculated over all targets are given in Table IV. To concentrate on targets that do have a domain boundary, we also computed these quantities after excluding all one-domain targets (34 target set) and also after excluding all targets that have a one-domain definition (17 target set). The results of these calculations are given in Table V. These tables show that human experts P0237 (Sternberg), P0536 (CBRC-DR), P0089 (KIAS), P0096 (CaspIta), and servers P0353 (Baker-Rosettadom) and P0421 (Baker-Robetta-Ginzu) are the top performing groups. (The rankings presented here are somewhat different from that presented at the December CASP meeting, mainly because the single domain predictions were handled differently. In the December ranking, all one-domain predictions were excluded in the average. Unfortunately, this introduced a distortion for the multidefinition targets because the NDO scoring scheme does not otherwise explicitly consider the number of domains. For this report, we excluded different sets of targets, not individual predictions, for ranking.)

The CBRC-DR group submitted predictions for all targets and recorded the highest average NDO score among

TABLE IV. Prediction Scores of Each Group for All 63 Targets

Prediction group	Np[a]	NDO score Mean	NDO score Rank	*Z*-score Mean	*Z*-score Rank	Top-1[b] Count	Top-1[b] Rank	Top-score[c] Count	Top-score[c] Rank
P0018	8	86.11	10	0.00	12	3	17	6	17
P0061	1	58.39	22	−1.30	22	0	20	0	21
P0063	56	60.94	20	−1.15	20	8	14	9	15
P0089	63	88.07	6	0.37	6	31	5	44	6
P0096	62	89.22	3	0.45	3	30	8	46	2
P0237	7	89.82	2	0.70	1	0	20	3	19
P0461	58	74.29	15	−0.35	15	18	11	22	12
P0536	63	90.73	1	0.51	2	36	1	50	1
P0590	61	87.17	8	0.33	7	31	5	40	8
P0667	56	65.57	17	−0.74	19	11	13	17	14
P0682	33	58.48	21	−1.26	21	3	17	3	19
Human	468	77.16		−0.22		171		240	
P0019	51	84.47	12	0.19	11	28	9	33	10
P0283	21	64.78	18	−0.60	17	1	19	4	18
P0289	2	63.66	19	−0.60	18	0	20	0	21
P0290	56	77.79	14	−0.21	14	8	14	22	12
P0309	61	82.40	13	0.00	12	16	12	31	11
P0353	63	88.74	5	0.39	4	33	3	45	3
P0381	60	86.16	9	0.25	9	28	9	40	8
P0421	63	88.76	4	0.39	4	34	2	45	3
P0435	63	85.08	11	0.20	10	31	5	43	7
P0436	63	87.69	7	0.30	8	32	4	45	3
P0638	40	72.03	16	−0.49	16	4	16	8	16
Server	543	80.14		−0.02		214		316	

[a]Number of predictions each group submitted.
[b]Number of times the prediction was best among all predictions for a given target.
[c]Number of times the prediction score was within 5% of the best score for a given target.

TABLE V. Prediction Scores of Each Group for Multidomain Targets

Prediction group	34 Target set Np[a]	34 Target set NDO score Mean	34 Target set NDO score Rank	34 Target set *Z*-score Mean	34 Target set *Z*-score Rank	17 Target set Np[a]	17 Target set NDO score Mean	17 Target set NDO score Rank	17 Target set *Z*-score Mean	17 Target set *Z*-score Rank
P0018	4	84.85	2	0.14	10	2	70.15	7	−0.41	16
P0061	0[b]	—	—	—	—	—	—	—	—	—
P0063	32	50.99	19	−1.20	20	16	51.49	19	−0.69	19
P0089	34	82.24	7	0.35	6	17	73.22	5	0.40	5
P0096	33	82.68	6	0.47	3	17	69.08	8	0.21	7
P0237	7	89.82	1	0.70	1	4	86.99	1	0.82	1
P0461	33	69.98	15	−0.19	15	17	67.39	10	0.19	8
P0536	34	84.60	3	0.50	2	17	75.01	4	0.49	4
P0590	33	80.25	8	0.28	7	17	69.02	9	0.16	9
P0667	32	57.36	18	−0.77	18	16	53.53	17	−0.59	18
P0682	22	50.86	20	−1.21	21	12	46.85	20	−1.06	21
Human	264	73.36		−0.09		135	66.27		−0.05	
P0019	29	75.86	13	0.10	12	15	66.51	11	0.04	10
P0283	16	65.30	17	−0.35	17	9	62.12	15	−0.22	14
P0289	1	39.00	21	−1.05	19	1	39.00	21	−1.05	20
P0290	31	72.62	14	−0.07	14	16	60.63	16	−0.31	15
P0309	34	77.41	12	0.08	13	17	63.68	13	−0.17	13
P0353	34	83.89	4	0.45	4	17	80.84	2	0.71	2
P0381	31	79.93	9	0.21	8	15	70.62	6	0.22	6
P0421	34	83.85	5	0.44	5	17	78.55	3	0.63	3
P0435	34	77.54	11	0.12	11	17	62.17	14	−0.14	11
P0436	34	79.71	10	0.17	9	17	64.20	12	−0.16	12
P0638	22	65.95	16	−0.32	16	10	52.19	18	−0.46	17
Server	300	72.82		−0.02		151	64.83		−0.05	

[a]Number of predictions each group submitted.
[b]Group P0061 predicted only one target T0197 (and T0202, which was canceled), which is a one-domain protein.

TABLE VI. Benchmark Scores

	All 63 targets				34 Target set		17 Target set	
Programs	Mean NDO score	Top-1 count	Mean *Z*-score	Top-score count	Mean NDO score	Mean *Z*-score	Mean NDO score	Mean *Z*-score
C1D	87.55	33	0.33	46	78.69	0.17	57.92	−0.49
CMD	74.04	20	−0.39	29	64.88	−0.53	61.36	−0.20
Domain Parser	93.71	37	0.70	54	90.10	0.85	82.84	0.98
PDP	94.54	39	0.78	54	93.29	1.08	90.72	1.54
PUU	94.69	39	0.78	56	91.67	0.99	89.05	1.40

all predicting groups. Fifty out of 63 predictions scored within 5% of the best, including 36 times when they were the best. They had the second best average *Z*-score. Sternberg group submitted predictions for only seven multidomain targets. Even though none of their predictions was best for any given target, the average *Z*-score was the best. If only the seven targets were used, the average NDO score was also the best (89.8 for P0237 vs. 88.9 for the next best by P0089). Also, for the 34- and 17-target sets, this group was rank 1 by both the average NDO and the average *Z*-scores. On the server side, the predictions from Baker group's two servers, P0353 and P0421, were the best whether all targets were considered or only the multidomain targets were considered. On multidomain targets, they were comparable or better than all human experts except Sternberg group.

On average, servers did slightly better than human experts when all targets were considered (Table IV), but slightly worse if only the multidomain targets were considered (Table V). On the other hand, servers had more top-1 and top-score counts than human experts for both all targets and multidomain only categories. This is true even after adjusting for the fact that servers made more predictions than humans.

Benchmark Calculations

To obtain upper and lower benchmark scores, we carried out similar calculations using the results from three domain parsing programs, PDP,[19] PUU,[20] and DomainParser,[21] as well as those from two controls, C1D and CMD. The domain parsing programs parse a known protein structure into domains; they are not domain prediction programs. The three programs chosen were the three that were most readily available to us at the time of the initial assessment. C1D predicts that all the targets are one-domain proteins. CMD cuts the target into two equal-sized domains if it has 100 to 200 residues and into three if it has more than 200 residues. The *Z*-score for these programs and controls were obtained by using the same mean and standard deviation values used to calculate the *Z*-scores for the predicted models for each target. The top-1 and top-score frequencies are the number of times a score was greater than the highest or within 5% of the highest prediction score, respectively, for each given target.

Table VI includes the average NDO score, top-1 counts, average *Z*-score and the top-score counts for these five programs. PUU and PDP had slightly better performance than DomainParser, but all three had greater than 90% NDO score average. This indicated that the three programs largely agreed with our manual domain definitions, and that the NDO scoring scheme does recognize similar domain parsing schemes as being similar. The average NDO score of C1D was better than seven human expert predictions and eight server predictions if all targets were included. Even when one-domain targets are excluded, C1D was better than four humans and six servers. This was partly because nearly half of the latter set of targets have one-domain as a domain definition, for which C1D gets the full score. The CMD, which cuts domains based only on the length of the sequence, is much poorer. Still, when only multidomain targets were considered, there were three human predictions and one server prediction that had average NDO score worse than CMD.

DISCUSSION

The NDO Scoring Scheme

The NDO scoring scheme gives a new quantitative scale for assessing domain boundary prediction. Compared to other existing measures, the chief distinguishing characteristics of the new scheme are the simplicity of its score and the lack of ad hoc parameters. For instance, if the predicted and defined boundaries are separated by n residues, and if there is no linker between the domains, the penalty is $2n$ (n for one domain being too large and another n for the other domain being too small). If there is a linker of length m between two domains and the predicted boundary falls within the linker, the penalty is $m/2$. If a two-domain target is predicted to be one domain, the penalty is equal to the number of residues in the smaller domain. The over- and underpredictions are treated in completely symmetric manner. In fact, if the role of the target and prediction are reversed, the raw score does not change. One possible drawback is that the score is measured relative to the perfect score by dividing the raw score, which is in units of the number of residues, by the perfect score. This was done to compare a prediction against multiple domain definitions. However, this also means that the same n-residue inaccuracy in the positioning of the boundary becomes a large or small NDO score depending, respectively, on whether the protein is small or large.

A sense of the NDO scoring scale of the prediction quality can be obtained by examining individual cases. Figure 6 shows a sample case of predictions for the target T0268, a two-domain protein without alternate definition.

For another example, consider two predictions for T0248, which is a three-domain protein (Fig. 1). As mentioned in the Results section, the top-scoring prediction [P0461, Fig. 1(B)] is also a three-domain but has the d2–d3 domain boundary in the middle of D3. Baker-Rosettadom (P0353), on the other hand, predicted a two-domain for this target with an accurate D1–D2 boundary, but one domain for both D2 and D3. Their NDO score was 54.35, compared to 61.07 for P0461. One has the virtue of not cutting in the middle of a domain but the other has the correct number of domains [see also Fig. 6(C) and (D)]. The NDO score provides a quantitative measure for ranking the two predictions, in such a manner that the ranking will change depending on how much d2 extends into D3 in the three-domain prediction. However, because it does not have an adjustable parameter, one cannot "fine tune" the scale to better fit our heuristic notion of the quality of prediction. In the future, it might be desirable to introduce a tunable parameter, for example, a weight between under- and overprediction or a parameter that will modify the current linear dependence of the overlap penalty on the number of residues.

We used the NDO scores exclusively for all the quantitative assessment of the domain predictions presented here and must point out that the predictor rankings we provide necessarily depend on the scoring scheme we used. To obtain a sense of the relation between NDO scoring and the scoring schemes that others have used, we computed the *Z*-scores for all predicting groups using three other scores besides the NDO scores. These are the average number of targets for which a group predicted the correct number of domains (CorrectNdom) and the specificity and the sensitivity of the domain boundary prediction. The latter two are defined as the average over all targets of TP/Npred and TP/Ndef, respectively, where TP is the number of predicted domain boundaries that are within ± 10 residues from a defined boundary and Npred and Ndef are the number of predicted and defined boundaries, respectively, for a given target. The raw scores, *Z*-scores and the rankings using these four scoring schemes are given in the Supplementary Table III. The targets used are the 17 targets that do not have a one-domain definition. The 17 target set is preferable for this purpose over the full 63 targets because the presence of one-domain targets severely distorts all three non-NDO scoring schemes. We find that there is generally a high correlation between the NDO *Z*-score rankings and the rankings based on the average of the three other *Z*-scores. For example, the four groups that score highest by the NDO *Z*-score (groups 237, 353, 421, and 536) are also the top four groups according to the average *Z*-score of these three other scores. The high correlation between the NDO *Z*-scores and the *Z*-scores of other three scoring schemes can also be seen in Figure 7.

The Ranking Issues

To evaluate the overall performance of the prediction groups, we computed the average NDO and *Z*-scores across the targets for each prediction group. We ignored nonsubmissions for this average. We are aware that submissions-only averages have their pitfalls: If a group makes predictions for only the easy targets, that group's average will be higher than that of another group who submits predictions for all targets. On the other hand, to include nonsubmissions, one must assign an arbitrary score to these instances. For those who submitted for only a small number of targets, the averages computed this way would not be useful. In the present study, we could spot the excellent performance of the Sternberg group (P0237, Tables IV and V) because we used the submissions-only averages. We verified that this group indeed performed well by noting (1) that the targets that this group selected to submit are multidomain, not particularly easy targets, and (2) that, in a head-to-head comparison using only these targets, this group does have the best average.

The high ranking in terms of the *Z*-score also distinguishes group P0237. This is in contrast to group P0018, who also submitted for only a few targets and ranked second in average NDO score for the multidomain targets (Table V). Two of its four predictions are single-domain and the scores are based on the single-domain definition. The ranking by the *Z*-score gives a better discrimination is such a case. Thus, the NDO and *Z*-scores are complementary and both useful.

Generally, the presence of single domain proteins hinders evaluation because it inflates the NDO scores. For a target with one- or two-domain definition, for example, the smart prediction is to predict one-domain, but the presence of such prediction hinders the detection of good two-domain predictions. For this year's set of targets, this problem is significant because all but 17 of the 63 targets have a single-domain definition. We have tried to avoid this problem by computing averages over only the multidomain targets and over only those that do not have a one-domain definition.

Methods Used by Top-Scoring Groups

Different groups used rather different methods to achieve good results. The Sternberg group (P0237) used the system named Phyre (private communication). It generates a 3D model from common structural regions of a number of templates obtained by an ensemble of fold recognition algorithms. The start and end points of this model defined the domain boundary. Any remaining section of the target sequence longer than 20 amino acids was resubmitted to the Phyre system for a recursive detection of additional domains.

The CBRC-DR group (P0536) (http://predictioncenter.llnl.gov/casp6/abstracts/abstracct.html) first searched possible coil regions using threading methods and then used the domain linker index[10] to select possible interdomain linkers from the coil regions identified in the first step.

The KIAS group (P0089)[13] used trained neural network to predict domain boundary from protein sequence information.

The CaspIta group (P0096) (http://predictioncenter.llnl.gov/casp6/abstracts/abstracct.html) used three different methods. Easy ones are identified using CDD.[22] Some domains were identified by the presence of particular folds as recognized by fold recognition. They also used the amino acid extension of PRIMEX,[23] which uses information on correlated sequence pattern to determine the likelihood that two fragments belong to the same domain.

Baker-Robetta-Ginzu (P0421) (http://predictioncenter-.llnl.gov/casp6/abstracts/abstracct.htm) used two basic steps. First, regions that were closely or remotely homologous to proteins of known structure were identified. Then multiple sequence alignment was used for the regions not identified in the first step. Baker-Rosettadom (P0353) (http://predictioncenter.llnl.gov/casp6/abstracts/abstracct.html) also uses two basic steps. It took the result of Ginzu for the regions that were homologous to sequences of known structures. For the remaining regions, it used Rosetta to obtain 200 3D models and a domain parsing process to obtain domains of these models. A consistent set of domains was obtained from these models.

Thus, most built one or more preliminary 3D models to determine at least some of the domains.

Comparison with CAFASP Results Is Difficult

It is difficult to compare the present evaluation results with the CAFASP evaluations for the following reasons. First, the lists of targets evaluated were different. CAFASP evaluated 58, lacking T0226, T0248, T0268, T0279, and T0280, mostly the difficult targets that we identified here. They also did not use alternate domain definitions and used only those predictions that did not have multiple segments. The predictors were different because they only evaluated automatic servers. The score used to measure the quality of predictions were also different. Because of these differences, CAFASP remains a valuable, independent resource.

Status of Domain Predictions

In general, good predictors did well with domain prediction. According to the average NDO scores given in Tables IV and V, they correctly assigned domains to more than 80% of the residues for the multidomain targets and more than 85% for all targets. The average NDO scores of top five groups were above 80. For all but four of the 63 targets, there was at least one prediction with an NDO score better than 90. However, some of the multidomain proteins and the proteins that are made of multiple segmented domains remain difficult to predict.

ACKNOWLEDGMENTS

We thank CASP organizers for maintaining the Web site and for providing the basic data and superb support. We thank Drs. Nick Grishin and S. Sri Krishna for suggesting domain parsing programs, and Drs. Liisa Holm and Jun-Tao Guo for providing us with the PUU and DomainParser programs.

REFERENCES

1. 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.
2. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
3. Gracy J, Argos P. Automated protein sequence database classification: II. Delineation of domain boundaries from sequence similarities. Bioinformatics 1998;14:174–187.
4. Guan X, Du L. Domain identification by clustering sequence alignments. Bioinformatics 1998;14:783–788.
5. George RA, Heringa J. Protein domain identification and improved sequence similarity searching using PSI-BLAST. Proteins 2002;48:672–681.
6. Rigden DJ. Use of covariance analysis for the prediction of structural domain boundaries from multiple protein sequence alignments. Protein Eng 2002;15:65–77.
7. Wheelan SJ, Marchler-Bauer A, Bryant SH. Domain size distributions can predict domain boundaries. Bioinformatics 2000;16:613–618.
8. Galzitskaya OV, Melnik BS. Prediction of protein domain boundaries from sequence alone. Protein Sci 2003;12:696–701.
9. Marsden RL, McGuffin LJ, Jones DT. Rapid protein domain assignment from amino acid sequence using predicted secondary structure. Protein Sci 2002;11:2814–2824.
10. Suyama M, Ohara O. DomCut: prediction of inter-domain linker regions in amino acid sequences. Bioinformatics 2003;19:673–674.
11. Nagarajan N, Yona G. Automatic prediction of protein domains from sequence information using a hybrid learning system. Bioinformatics 2004;20:1335–1360.
12. Liu J, Rost B. Sequence-based prediction of protein domains. Nucleic Acids Res 2004;32:3522–3530.
13. Sim J, Kim SY, Lee J. PPRODO: prediction of protein domain boundaries using neural networks. Proteins 2005;59:627–632.
14. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res 1999;27:260–262.
15. Ponting CP, Schultz J, Milpetz F, Bork P. SMART: identification and annotation of domains from signalling and extracellular protein sequences. Nucleic Acids Res 1999;27:229–232.
16. Tress M, Chin-Hsien T, Wang G, Ezkurdia I, López G, Valencia A, Lee BK, Dunbrack RL Jr. Domain definition and target classification for CASP6. Proteins 2005;Suppl 7:8–18.
17. DeLano WL. The PyMOL Molecular Graphics System (2002) on World Wide Web http://www.pymol.org.
18. Vincent JJ, Tai C-H, Sathyanarayana BK, Lee B. Assessment of CASP6 predictions for new and nearly new fold targets. Proteins 2005;Suppl 7:67–83.
19. Alexandrov N, Shindyalov I. PDP: protein domain parser. Bioinformatics 2003;19:429–430.
20. Holm L, Sander C. Parser for protein folding units. Proteins 1994;19:256–268.
21. Xu Y, Xu D, Gabow. N. Protein domain decomposition using a graph-theoretic approach. Bioinformatics 2000;16:1091–1104.
22. Marchler-Bauer A, Anderson JB, DeWeese-Scott C, et al. CDD: a curated Entrez database of conserved domain alignments. Nucleic Acids Res 2003;31:383–387.
23. Lexa M, Valle G. PRIMEX: rapid identification of oligonucleotide matches in whole genomes. Bioinformatics 2003;19:2486–2488.