PROTEINS: Structure, Function, and Genetics, Suppl. 1:92–104 (1997)

Competitive Assessment of Protein Fold Recognition and Alignment Accuracy

Michael Levitt*
Department of Structural Biology, Stanford University School of Medicine, Stanford, California

ABSTRACT **The predictions made for fold recognition and modeling accuracy at the 1996 Critical Assessment of Structure Prediction meeting (CASP2) were assessed to discover which groups did best. With 32 groups making a total of 369 predictions, it was necessary to use simple criteria for distinguishing between the entries. By focusing on the predictors' ability to use the sequence of the unknown target structure to recognize the target fold from a database of known folds and also on the quality of the model judged by the accuracy of the predicted alignment, it is easy to determine the best predictions for a given target. Assessing overall performance of the predictors on all the targets is much more difficult and use was made of weighted averages of fold recognition and alignment accuracy with and without normalization for target difficulty. By plotting these results in two dimensions the winning groups stand out, allowing readers to focus their attention on the most promising methods. When the present results are compared with the results of the earlier CASP1 meeting, held in 1994, it is clear that threading predictions have progressed dramatically. For this assessor, the strongest lesson learned is that subjectivity is pervasive and affects us all. It is abundantly clear that the blind predictions made at CASP are essential if progress is to be made in predicting protein structure. Proteins, Suppl. 1:92–104, 1997.** © 1998 Wiley-Liss, Inc.

Key words: protein folding; fold recognition; threading; alignment accuracy; CASP; Asilomar

INTRODUCTION

What is the role of the assessor? As a judge at any competitive event, one is expected to pick those entries considered best. This needs to be done in a way that is based on clear objective criteria and eliminates subjective choices. Picking the best entries is not a problem when there is a small number of entrants, as they can all be winners. In the context of the Critical Assessment of Structure Prediction (CASP) meeting, the purpose of the assessment is to focus attention on those predictors that are doing best: a participant at the meeting should go away knowing which papers to read. There is also tremendous pressure to have this assessment done as quickly as possible.

The task facing the assessment of the threading predictions at the 1996 CASP2 meeting was particularly difficult. First, there was a very large number of entries, with over 32 groups participating. Second, there was a massive amount of data submitted for each entry, which consisted of a list of recognized folds (PDB identifiers) together with one or more predicted alignments to a particular target. There were 15 different targets on which predictions were submitted by 32 different groups, giving a total of 369 individual predictions. Third, evaluation of threading is difficult, involving, as it does, two separate criteria—fold recognition and model accuracy. Both of these can be considered from the point of view of the experimental biologist who comes to a computational biologist with a new sequence that he has not been able to match using conventional sequence database searches. Two commonly asked questions will be: 1) Does this sequence match a protein sequence for which there is a known three-dimensional structure? and 2) If it matches, can one align enough of the target sequence to the known fold to build a useful three-dimensional model for the new sequence?

Faced with these difficulties, two basic decisions were made: 1) All assessment of threading would rely entirely on the evaluation provided by Marchler-Bauer and Bryant.[1] In the context of CASP2, evaluation is defined as the process of data reduction whereby the large body of raw data on each prediction is reduced to a few evaluation indices, which were carefully chosen and agreed upon before the meeting. Assessment involves judging the relative performance of the predictors based on how well they score in terms of these evaluation indices. While the raw data were made available to me as assessor, I chose not to use them except for cross-checking

Contract grant sponsor: NIH; Contract grant number: GM41455; Contract grant sponsor: DOE; Contract grant number: DE-FG03-95ER62135.
*Correspondence to: Dr. Michael Levitt, Department of Structural Biology, Stanford University School of Medicine, Stanford, CA 94305.
E-mail: michael.levitt@stanford.edu
Received 14 August 1997; Accepted 25 August 1997

purposes. This decision was made after an attempt at independent evaluation and was based on the exceptional job done by the threading evaluators and the desire to separate assessment from evaluation. 2) All assessment would be blind, in that the assessor should not know which group had made which prediction. This is essential to combat the subjectivity that is an insidious enemy of fair assessment.

Choice of speakers at the meeting had to be made at very short notice. Therefore, it was based on performance on a single target as measured by the number of correctly aligned residues to a correctly recognized fold (the evaluation index ACrct, see below). Choice of predictors to write articles in this issue was made after more than three months of additional analysis. It was based on sustained performance on different targets as measured by many different evaluation indices. This is a much more difficult task and is beset by problems associated with the variation in target difficulty, the choice of suitable evaluation indices, and the manner in which different indices are combined. Nevertheless, I believe that the choice of speakers and contributors of articles to this issue does include the groups who did best. Some of the decisions were close and small changes could have affected a few of the choices. This instability is inherent in any attempt to choose winners, be it the Olympic Games or the CASP competition. The alternatives of letting all predictors be winners or choosing winners randomly would certainly be less satisfactory.

METHODS AND RESULTS

Choice of Evaluation Criteria

The evaluators provided this assessor with a large number of different assessment criteria. The choice of the best criteria was based on the two perceived goals of threading: 1) to recognize the correct fold and 2) to predict the alignment accurately. The CASP1 assessment had indicated that correct alignment was a better discriminator of good predictions than fold recognition alone. It is important to realize that reliable and correct fold recognition is important, as without it there can be no assessment of alignment accuracy. Here we use the word "threading" loosely to cover the general task of recognizing a known fold that matches the target sequence, together with prediction of how this sequence is aligned to the recognized fold.

Based on these arguments and the desire to have as few criteria as possible, focus was placed on threading specificity, TSpc and number of correctly alignment residues, ACrct (also referred to as the alignment accuracy). For each database protein, i, the predictor assigned a normalized score or probability, $AWgt_i$ (between 0 and 1) that this protein has the same fold as the target sequence. They also predicted how the target sequence is aligned to this database protein (set of equivalence pairs).

As soon as the structures of the targets became available, the evaluators used three structure comparison methods, Dali[2,3], SSAP[4,5], and VAST[6] to find database proteins that do indeed have the same fold as the target. These same methods also produced a structural alignment that equivalences some residues of the target sequence with residues in the fold recognized. The evaluators counted how many of the predicted equivalence pairs were identified by each structural alignment method, m, to give the number of correctly aligned residues, $ACrct_{im}$. This count, an extensive measure, was favored since a correct alignment will be better if it involves more residues, allowing one to build a better homology model.

For each target, the evaluators then used $AWgt_i$ and $ACrct_{im}$ to average over the recognized folds and calculate Conf, TSpc, and ACrct for the particular prediction. The confidence in the prediction, Conf, is calculated as

$$\mathrm{Conf} = 100\left(\sum_i \mathrm{AWgt}_i\right),$$

where the sum is over all database proteins. Assigning a non-zero weight to none of the database proteins (the "NONE" fold) will reduce Conf from its maximum value of 100. This will reduce the weight of the prediction in the averaging of results over different targets. Thus, a bad prediction made with low confidence will have a smaller effect on sustained performance than the same prediction made with high confidence.

The threading specificity, $TSpc_m$ for structure comparison method m, is calculated as

$$\mathrm{TSpc}_m = 100 * \left(\sum_I \mathrm{AWgt}_i * \mathrm{SCWgt}_{im}\right),$$

where the sum is over all database proteins. The structure comparison weight, $SCWgt_{im}$, varies between 0 and 1, depending on how closely the database fold used in the i-th prediction is judged to match the target structure. For the VAST method $SCWgt_{im} = 1$ if the fold is recognized and $= 0$ otherwise, whereas for Dali and SSAP, SCWgt varies continuously between 1 and 0.

The number of correctly aligned residues or alignment accuracy, $ACrct_m$, is calculated as

$$\mathrm{ACrct}_m = \left(\sum_i \mathrm{ACrct}_{im} * \mathrm{AWgt}_i * \mathrm{SCWgt}_{im}\right) \Big/ \left(\sum_i \mathrm{AWgt}_i * \mathrm{SCWgt}_{im}\right)$$

where $SCWgti_m$, $ACrct_{im}$, and $AWgt_i$ are defined above. The appearance of $SCWgt_{im}$ in the denominator ensures that $ACrct_m$ averages over only those models that are based on correctly recognized folds. This means that if the alignment is wrong, the overall alignment accuracy will be higher if the fold is incorrect. While this does not seem fair, it will be reflected in a less good value of the threading specificity, $TSpc_m$.

Is this choice of TSpc and ACrct as primary assessment indices justified? A study of the correlation between the nine different indices tabulated by Marchler-Bauer and Bryant[1] showed that four of the indices concerned with alignment and contact sensitivity and specificity (ASpc, ASns, ACSpc, and ACSns) are all correlated to one another with a correlation coefficient of at least 0.79. In addition, these four indices are also correlated to the threading specificity, TSpc, by at least 0.66. This leaves six indices: Conf, TSpc, ACrct, ARms, Shft, and Covr, the first three of which have been described above.

Two of the additional indices, Modeling Accuracy (ARms) and Alignment Shift (Shft), both measure the error in the predicted models. ARms is the root mean square deviation of the target structure and the recognized fold, calculated using the CA coordinates of corresponding residues in the predicted alignment of target sequence and recognized fold. Shft measures the error in the predicted alignment of the target sequence to the recognized fold relative to the actual structural alignment. Here more attention is focused on ARms, as it is a more conventional measure of accuracy in protein modeling. ARms has a major advantage over all other indices measuring model accuracy in that it does not depend on structural comparison: the predicted alignment of database fold to target is simply used to calculate an RMS deviation. The RMS deviation is the only index that is shared by all four sections of CASP, allowing cross-comparison. The third index, alignment coverage, Covr, measures the percentage agreement between the predicted alignment and that found by structure comparison; it shows how much of the target structure is predicted. In the tables for each target, ARms, Shft, and Covr are averaged over recognized folds weighting by $AWgt_i * SCWgt_{im}$ (in the same way as for ACrct, above).

In calculating the sustained performance, the selected evaluation indices are averaged over the different groups or different targets. This averaging is weighted by Conf, the prediction confidence, so that a prediction made with more confidence scores better than one made with less confidence. As explained by Marchler-Bauer and Bryant[1], Conf is less than 100 when the predictor has one of his choices as "NONE," that is, that the target fold is not in the database. Specifically, the average of property A is calculated as:

$$A = \left(\sum_j (A_j * Conf_j)\right) \Big/ \left(\sum_j Conf_j\right)$$

where the summation, j, is either over groups for each target or over targets for each group. The denominator is related to the effective number of entries calculated as

$$Nent = \left(\sum_j Conf_j\right) \Big/ 100$$

As Σ Conf never exceeds 100, Nent is less than or equal to the number of predictions for each target by each group. When a group submitted a number of independent predictions for a particular target, the total Conf value for that target was normalized to not exceed 100.

For the easier targets, the values of each property, A, will be better and this will dominate the average. Here, I allow for this by calculating a normalized A' for each target as follows:

$$A'_j = A_j \Big/ \left(\sum_k A_k * 0.01 * Conf_k\right)$$

where the summation k is over all the predictions for target, j. Thus, each property is divided by the mean value of the property for the particular target.

Some Technical Issues

Although almost all processing of entries was done for me by the evaluators and others associated with the massive task of data collection and dissemination, there were some minor issues that required attention. One was the removal of duplicate entries in which a predictor had submitted the same entry more than once. This is an easy task but it does raise the issue of how to deal with multiple entries. Some groups submitted multiple independent entries and these were averaged so that no advantage was obtained. More difficult is the submission of multiple entries by different groups that both include the same individual predictor. In some cases, this reflected independent predictions of different members of a particular laboratory and all included the name of the group leader. In other cases, the individual predictor participated in different independent groups. Here, no attempt is made to allow for this, but it did influence the choice of speakers in one instance.

Averaging is also complicated by the fact that in the tables provided by the evaluators a missing entry is given a value of 0.0. This works when a zero value

is the worst possible score—for example, if no fold is recognized, the threading specificity, TSpc, is zero, and if no residues are correctly aligned, the number of correct alignments, ACrct, is zero. It fails for measures like modeling accuracy, ARms, and alignment shift, Shft, which score zero for perfect predictions. In calculating averages, all entries which failed to recognize any fold (TSpc = 0), were excluded from averages of ARms and Shft. These entries were, however, included in averages of all other evaluation indices for which the failure to recognize the fold scores zero.

The predictions for the target T0002 were not available before the choice of speakers and the results tabulated at the meeting were misleading as this target has two domains, only one of which was a proper threading target. The other domain was, in fact, a target for homology modeling: many threading entries inappropriately predicted a fold and alignment for this domain. The correct evaluations for this target were not made available to me until the beginning of June but they are included in the results presented here.

Consistency of Structural Alignment

All the threading predictions were evaluated by comparison to structural alignments produced by well-established methods such as Dali, SSAP, and VAST. If these methods fail to recognize a database protein as being structurally similar, or do not give the best possible structural alignment, the threading predictions for that target will be wrongly scored. Structural alignment is beset by three potential shortcomings: 1) unlike sequence alignment, structural alignment is not globally convergent and the results are sensitive to the heuristic search strategy; 2) judging fold-recognition involves a statistical estimate of the significance of the match, which is a complicated issue even for sequence alignment; and 3) any structural alignment can always be trimmed to have fewer matches with a lower RMS deviation and it is not clear when to stop this procedure. Because of these difficulties, concern has been expressed as to whether there really is a best structural alignment.[7,8]

The three methods of structural alignment used by the threading evaluators are all well established and involve different algorithmic approaches. Figure 1 shows a Venn diagram indicating the extent to which these methods agree. Agreement between structural comparison methods is defined very loosely: for a particular entry, two methods are deemed to agree if they both find that at least one of the database folds predicted to fold like the target is in fact a significant structural match. All three methods agree for 84 predictions, whereas one or two of the three methods finds a match in an additional 78 cases. Clearly, there is not perfect agreement between these methods.

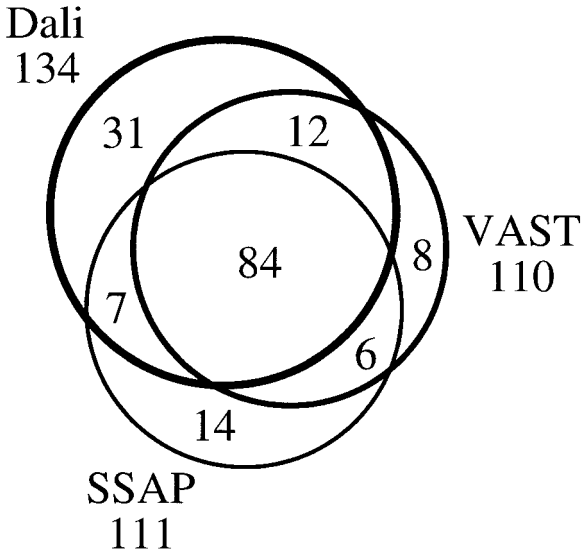


Fig. 1. A Venn diagram that counts the predictions with a non-zero TSpc score for each of the three methods (Dali, SSAP, and VAST) used to detect similarity between the actual structure and a database of known folds. A zero TSpc value indicates that none of the database folds predicted to be similar to the target was judged structurally similar by the particular method. This could occur either because the predictions are all wrong or because the method used to detect structural similarity fails. There are a total of 265 predictions (completely excluding target T0002, for which there were no SSAP results). At least one of the three methods gets a non-zero TSpc score for 162 of the predictions (the sum of counts for all seven regions of the Venn diagram), whereas all three methods get non-zero TSpc for only 84 predictions. There is slightly better agreement between the Dali and VAST methods, with 96 in common out of a total of 148 predictions.

Here, we use the agreement between the structural comparison methods and the predictions in an attempt to further compare the methods. For two of the methods, Dali and VAST, the predictions score better, with overall average threading specificity that is over 60% higher than for the SSAP method (TSpc = 57.7% and 64.9%, relative to 36.0%). The alignment accuracy for the Dali and VAST is also 34% higher than for SSAP (ACrct 26.7 and 26.3, relative to 19.6). The SSAP results were not used in this assessment, for three reasons: 1) they fit the predictions less well, 2) the results were not completed in time for the original assessment, and 3) the best fold matches (in terms of number of residues matched, SCLen, and the RMS deviation of these residues, SCRms) were always found by Dali or VAST. Initially, it was not clear why SSAP scores less well than Dali or VAST. It was subsequently indicated that the version of SSAP used for the evaluation was the fast version, which only compares secondary structure environments.[9] For distant relatives, the alignments are considered to be unreliable (Orengo, personal communication). Given this limitation, it is impressive how much agreement there is between the three methods. All predictions were scored by averaging the results for the Dali and

VAST methods for the 96 cases where they both found a fold (TSpc >0). In the remaining 66 cases, the score of either method was used.

Although there is good overall agreement in the average values of threading specificity and alignment accuracy for the Dali and VAST analyses, a plot of the Dali and VAST TSpc values shows that there is a low level of correlation (Fig. 2a). There are predictions for which the VAST score is good (above 50%), whereas the Dali score is poor (below 25%). The opposite situation also occurs, with six cases for which the Dali score is good (above 50%) whereas the VAST score is poor (below 25%). The corresponding plot for the ACrct values (Fig. 2b) shows that these values are much more similar for the Dali and VAST methods. In certain cases the Dali and VAST alignment accuracy values for a particular model differ significantly and averaging, as is done here, gives a lower score than the best value. Nevertheless, such averaging is considered to be more robust.

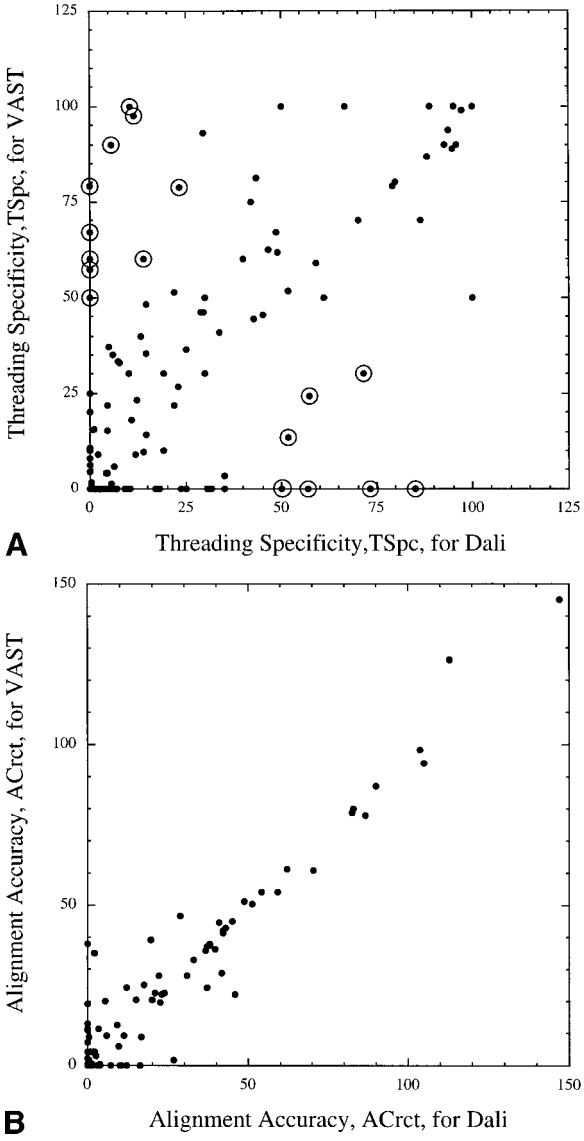


Fig. 2. (**A**) The individual TSpc scores with the Dali and VAST structural comparison methods are only weakly correlated. In particular, there are many predictions for which the VAST TSpc score is more significant (larger) than that from Dali. The circled points are those for which the Dali and VAST TSpc scores differ appreciably. (**B**) The individual ACrct scores are much more correlated for the Dali and VAST methods.

Overall Results for Each Target

Table I shows the overall results for each of the 15 targets. The average effective number of predictions (Nent) is 14.6. Eight of the targets had structures that were not similar to any known fold and no threading predictions were possible. Seven of the targets were found to be similar to database folds but the ability to recognize these folds and predict the correct alignment differs from target to target. There are three easy targets (T004, T0014, and T0031) whose average TSpc and ACrct values exceed 45% and 17.5, respectively. The other four recognizable targets (T0002, T0020, T0022, and T0038) are harder, with lower average TSpc values between 6.0 and 21.5% and ACrct values between 0.1 and 3.6. Note that this definition of easy and hard depends solely on how well the CASP2 predictors succeeded in their predictions. The average number of entries for the hard targets (16.4) is almost as high as for the easy targets (22.2). This same trend is seen for the predictor's confidence, Conf, which averages 68.8 and 76.0 for these two sets of targets, respectively. Clearly, the predictors were prepared to tackle the more difficult targets with confidence (Marchler-Bauer et al.[11] consider the issue of false positives, where a database fold is recognized when none should be).

Results on Individual Targets

Tables II to VIII present the results for each prediction for the seven targets for which fold recognition is possible. These tables do not present all results for each target; I omit predictions that have an ACrct value less than a cutoff determined either by the number or the quality of the predictions, depending on whether the target is easy or difficult. Names of all predictors are used in the tables, while the figures use the name of the first author as the group name. In two cases, the group name used here differs from the "official" group name: the di Francesco group is known as the Munson group and the Karplus group is known as the Haussler group. The first target, T0002 (Table II) was not solved in time for the results to be used to select speakers. The confusion about which domain to predict also meant that the results needed to be specially processed. In spite of the difficulty of this target, Murzin and Bateman do well, with perfect fold recognition and a good model with a significant number of correctly aligned residues. The value of ARms of 2.9 Å is

TABLE I. Mean[1] Results for Each Target

Target	Length	Ngrp[2]	Nent[3]	Conf	TSpc	ACrct	ARms	Shft	Covr	Npred[2]
Ordered by target number										
T0002[4,5]	514	19	16.5	71	48.2	47.9	12.9	7.8	34.8	25
T0002[4,5]	186	19	15.1	65	6.0	1.7	9.3	3.3	4.8	25
T0004[4]	84	24	22.5	75	45.5	17.9	5.8	1.6	57.0	37
T0005	269	10	8.7	86	0.0	0.0	0.0	0.0	0.0	12
T0010	456	15	8.7	54	0.6	0.0	28.0	0.0	2.0	21
T0011	220	15	10.2	73	0.0	0.0	0.0	0.0	0.0	18
T0014[4]	252	24	21.6	67	56.2	17.5	11.9	23.3	64.5	34
T0016	312	15	13.2	69	0.8	0.3	17.6	14.9	5.4	20
T0020[4]	310	23	19.6	63	14.5	0.9	18.9	44.4	19.2	31
T0022[4]	591	14	9.4	62	17.0	0.1	23.1	129.5	15.5	16
T0030	66	19	12.9	61	0.0	0.0	0.0	0.0	0.0	28
T0031[4]	242	24	22.5	86	72.7	51.3	10.7	3.6	67.0	29
T0032	98	18	12.3	68	0.0	0.0	0.0	0.0	0.0	22
T0037	108	17	12.3	64	0.0	0.0	12.4	0.0	5.4	21
T0038[4]	152	23	21.3	85	21.5	3.6	14.9	19.1	32.7	26
T0042	78	12	9.2	76	0.4	0.0	18.1	0.0	3.3	12
TSpc >1, Ordered by decreasing ACrct[6]										
T0031[4]	242	24	22.5	86	72.7	51.3	10.7	3.6	67.0	29
T0004[4]	84	24	22.5	75	45.5	17.9	5.8	1.6	57.0	37
T0014[4]	252	24	21.6	67	56.2	17.5	11.9	23.3	64.5	34
T0038[4]	152	23	21.3	85	21.5	3.6	14.9	19.1	32.7	26
T0002[4]	186	19	15.1	65	6.0	1.7	9.3	3.3	4.8	25
T0020[4]	310	23	19.6	63	14.5	0.9	18.9	44.4	19.2	31
T0022[4]	591	14	9.4	62	17.0	0.1	23.1	129.5	15.5	16

[1]The mean values of each index is obtained by averaging over all the predictions for a given target. In doing this averaging, care was taken to weight the quantities by the confidence the predictors placed on their prediction (weight = Conf/100).
[2]Ngrp is the number groups predicting this target. Npred is the total number of predictions made for the target. Because some groups submitted more than one prediction, Npred can be greater than Ngrp.
[3]Nent is the effective number of entries calculated as the sum of (Conf/100) over all predictions for each target value. Nent is less than Ngrp where part of the bet was placed on NONE so that the confidence (Conf) in the prediction is less than 100.
[4]These targets are all similar to at least one database fold as follows:

Target	T0002	T0004	T0014	T0020	T0022	T0031	T0038
Best Fold	1PSD:A	1CSP	1NAL:1	8ABP	1TCA	3EST	1BYH
SCLen	64	61	194	205	119	185	119
SCRms (Å)	1.88	2.19	2.84	4.96	3.14	2.31	2.92

where SCRms is the Cα RMS deviation and SCLen is the number of matching residues in the Dali structural comparison of the target and the best fold (the Dali result was chosen as it always had the smallest value of SCRms/SCLen).
[5]T0002 has two domains, only one of which is a threading target. Many entries ignored instructions to this effect and included the other domain, which has obvious sequence similarity to a database protein, in their predictions. These entries score artificially well and we focus on the difficult domain of T0002.
[6]The targets with the highest overall TSpc scores are shown in the lower part of the table, sorted by the overall ACrct score.

TABLE II. Better Results[1] for Target T0002 Sorted by ACrct

Entry Code[2]	Conf	TSpc	ACrct	ARms	Shft	Covr	Predictors
T0002FR220	90	100.0	26.2	2.9	0.2	46.1	Murzin & Bateman
T0002FR160	100	0.0	2.5	11.9	6.0	24.6	Barton & Copley
T0002FR244	67	1.1	0.0	14.1	245.2	9.4	Jones

[1]In Tables II–VIII, we only present those predictions that scored well. This cutoff, which avoids having to list the less good predictions, was determined either by the number or the quality of the predictions, depending on whether the target is easy or difficult.
[2]The entry code is the identifier assigned to the particular CASP2 prediction entry.

particularly impressive for a prediction of 46% of the 186 residues in the "threading domain" of T0002. The best prediction for this target is much better than all the other predictions.

T0004 (Table III) is a small protein (84 residues) that is well recognized and accurately aligned by many predictors. The best predictions correctly aligned more than half the residues and had ARms values below 3.5 Å for more than 80% of the residues. T0014 (Table IV) is a much larger protein (252 residues), which is also well recognized, with more than half the entries scoring perfect fold recognition (TSpc of 100). Correct alignments were more difficult: only three predictions correctly aligned more

TABLE III. Better Results for Target T0004 Sorted by ACrct

Entry Code	Conf	TSpc	ACrct	ARms	Shft	Covr	Predictors
T0004FR292	100	100.0	43.0	3.1	0.3	89.2	Alexandrov & Zimmer
T0004FR281	50	100.0	42.0	3.2	0.3	85.9	●● Kim & Dubchak*
T0004FR302	100	30.0	41.5	3.2	0.3	83.3	Solovyev
T0004FR148	100	100.0	38.0	4.8	0.3	73.8	Luethy, Alexandrov, Bass & Solovyev
T0004FR207	100	98.1	37.8	3.4	0.5	80.3	Eisenberg, Weiss, Rice & Fischer
T0004FR93	20	59.1	37.0	3.7	0.3	80.3	Coulson
T0004FR278	100	100.0	36.1	3.6	0.4	80.9	Karplus, Sjölander, Barret, Cline, Haussler, Hughey, Holm & Sander
T0004FR94	100	83.3	35.1	6.0	0.9	77.1	Sippl, Floeckner, Domingues & Jaritz
T0004FR242	74	61.2	34.1	3.7	0.7	80.4	Jones
T0004FR75	20	4.2	33.0	3.4	0.4	83.3	Coulson
T0004FR111	100	100.0	33.0	4.6	0.8	82.1	Hubbard, Park & Reinhardt
T0004FR224	100	50.0	24.8	5.2	1.0	80.2	Lengauer, Mevissen, Thiele & Zimmer
T0004FR112	76	11.8	21.1	7.4	19.8	79.6	Elofsson
T0004FR65	10	10.1	19.0	7.6	1.7	61.9	Coulson
T0004FR213	100	58.5	10.2	3.9	1.2	89.7	Murzin & Bateman

*In Tables III to X we mark the group chosen to speak at the meeting by "●●."

TABLE IV. Better Results for Target T0014 Sorted by ACrct

Entry Code	Conf	TSpc	ACrct	ARms	Shft	Covr	Predictors
T0014FR173	80	100.0	88.5	5.9	3.8	86.1	●● Murzin & Bateman
T0014FR267	100	97.5	50.1	12.3	10.6	82.2	Hubbard, Park & Reinhardt
T0014FR131	100	80.0	42.7	9.0	7.5	75.7	Sternberg, Bates, Russell, Saqi & Sayle
T0014FR178	100	100.0	37.6	10.2	9.7	78.9	Di Francesco, Geetha, Garnier & Munson
T0014FR337	100	75.0	30.5	10.9	7.8	74.1	Sippl, Floeckner, Domingues & Jaritz
T0014FR192	100	79.3	29.4	9.4	7.4	79.9	Bryant, Hogue, Madej, Marchler-Bauer & Ohkawa
T0014FR209	100	78.2	29.2	13.2	9.0	69.0	Luethy, Alexandrov, Bass & Solovyev
T0014FR170	100	91.4	21.7	12.8	6.1	67.9	Valencia, Pazos, Olmea & Rost
T0014FR151	100	26.6	19.0	15.1	57.8	47.8	Taylor & Munro
T0014FR269	20	100.0	17.8	13.9	16.6	73.4	Moult, Milash, Braxenthaler, Pedersen & Samudrala
T0014FR243	50	62.3	12.7	11.4	28.7	72.0	Jones
T0014FR208	80	94.4	10.7	11.9	27.2	72.2	Eisenberg, Weiss, Rice & Fischer
T0014FR299	100	4.1	8.2	9.0	3.9	81.2	Lathrop, Rogers, Smith & White
T0014FR205	100	14.6	7.4	17.4	44.4	63.4	Torda, Huber, Dyer & Lu
T0014FR154	100	4.1	7.2	8.8	63.7	81.1	Barton & Copley

than 40 residues (16% of the chain) and of these, only the best prediction (with ACrct = 88.5) had a good ARms (5.9 Å). T0031 (Table V), another large protein (242 residues), which is the third member of the easy target set, was also easily recognized by most predictors. The alignment accuracy was remarkably high for the best prediction by the Sippl group, with 146 correctly aligned residues (59%) and an ARms of 4.2 Å!

The three remaining targets are much more difficult to predict accurately. For T0020 (Table VI), a large protein (310 residues), fold recognition is not easy, with an average TSpc of 14.5%. The best prediction, by the Eisenberg group, managed to correctly align 14 residues out of the 132 residues, but the ARms value of 15.3 Å A is very high. For T0022 (Table VII), a very large protein (591 residues), the best prediction (Rost) has a reasonable TSpc of 54.5% but only correctly predicts one position of the alignment. The ARms values are also very high (above 20 Å) except for the prediction by the Lengauer group, for which ARms = 13.7 Å. For T0038 (Table VIII), a medium-sized protein (152 residues), recognition is only good for the best prediction with TSpc of 75% for Murzin and Bateman. The four best predictions have ACrct greater than 12, but the ARms values are reasonable (<10 Å) for only two groups, Murzin and Bryant.

Sustained Performance of Predictors

It is clear from the presentation of the individual prediction results that there is a wide spread of scores and that some targets are easier than others. Here, I calculate averages over the seven targets for which there are recognizable folds (Table IX), and also calculate averages normalizing for target difficulty (Table X) using the average values for each target given in Table I. Another consideration in measuring sustained performance must be the number of targets for which predictions were made, and

TABLE V. Better Results for Target T0031 Sorted by ACrct

Entry Code	Conf	TSpc	ACrct	ARms	Shft	Covr	Predictors
T0031FR795	100	100.0	146.0	4.2	0.3	86.6	●● Sippl, Floeckner, Domingues & Jaritz
T0031FR830	100	100.0	119.5	8.2	1.8	85.3	Dixon & Thomas
T0031FR861	100	100.0	101.0	8.0	1.8	86.0	Karplus, Sjölander, Barret, Cline, Haussler, Hughey, Holm & Sander
T0031FR854	90	100.0	99.5	6.9	1.4	85.8	Alexandrov & Zimmer
T0031FR853	90	100.0	99.5	7.4	1.5	86.0	Lengauer, Mevissen, Thiele & Zimmer
T0031FR858	100	100.0	82.2	9.1	2.2	86.2	Luethy, Alexandrov, Bass & Solovyev
T0031FR839	50	93.8	81.5	7.8	1.9	86.1	Coulson
T0031FR837	100	100.0	80.7	7.3	1.6	84.8	Eisenberg, Weiss, Rice & Fischer
T0031FR855	100	100.0	65.4	9.2	1.4	72.5	Valencia, Pazos, Olmea & Rost
T0031FR509	100	100.0	61.5	13.1	5.6	85.7	Sternberg, Bates, Russell, Saqi & Sayle
T0031FR836	86	51.7	56.7	11.8	4.2	84.2	Jones
T0031FR840	100	100.0	54.0	12.1	5.0	84.9	Taylor & Munro
T0031FR856	90	100.0	50.9	10.7	4.0	85.9	Hubbard, Park & Reinhardt
T0031FR864	100	100.0	45.0	13.3	5.0	87.5	Solovyev
T0031FR829	100	100.0	37.7	13.2	5.1	80.2	Di Francesco, Geetha, Garnier & Munson
T0031FR874	100	70.0	23.0	16.9	11.9	78.1	Torda, Huber, Dyer & Lu
T0031FR541	100	22.0	20.2	9.4	5.1	90.3	Honig, Yang & Xiao
T0031FR846	48	33.3	7.1	14.0	2.7	26.0	Zhou & Abagyan

TABLE VI. Better Results for Target T0020 Sorted by ACrct

Entry Code	Conf	TSpc	ACrct	ARms	Shft	Covr	Predictors
T0020FR710	47	51.0	14.0	15.3	53.2	42.7	●● Eisenberg, Weiss, Rice & Fischer
T0020FR749	50	14.3	7.7	14.8	12.2	51.5	Bryant, Hogue, Madej, Marchler-Bauer & Ohkawa
T0020FR728	50	3.5	3.0	19.3	95.2	32.7	Jones
T0020FR767	16	36.6	2.0	20.8	34.4	41.4	Lathrop, Rogers, Smith & White
T0020FR764	16	9.8	1.8	19.9	45.4	39.8	Lathrop, Rogers, Smith & White
T0020FR537	100	14.3	1.1	18.7	56.3	39.4	Honig, Yang & Xiao
T0020FR759	16	20.3	0.6	14.8	14.8	38.7	Lathrop, Rogers, Smith & White
T0020FR634	100	37.4	0.6	21.0	60.0	36.5	Sanejouand
T0020FR761	16	13.2	0.5	14.9	17.1	34.5	Lathrop, Rogers, Smith & White

TABLE VII. Better Results for Target T0022 Sorted by ACrct

Entry Code	Conf	TSpc	ACrct	ARms	Shft	Covr	Predictors
T0022FR348	100	54.5	0.9	28.4	239.1	10.6	●● Rost
T0022FR435	41	10.7	0.3	21.3	80.1	10.6	Jones
T0022FR436	100	17.7	0.1	23.5	40.3	22.5	Bryant, Hogue, Madej, Marchler-Bauer & Ohkawa
T0022FR321	20	37.0	0.0	27.8	324.7	13.3	Sternberg, Bates, Russell, Saqi & Sayle
T0022FR412	11	30.7	0.0	25.0	60.1	30.5	Eisenberg, Weiss, Rice & Fischer
T0022FR409	50	24.9	0.0	26.5	139.7	16.9	Coulson
T0022FR439	60	20.3	0.0	13.7	35.5	45.9	Lengauer, Mevissen, Thiele & Zimmer
T0022FR398	100	20.0	0.0	24.3	180.8	16.6	Torda, Huber, Dyer & Lu
T0022FR385	100	12.5	0.0	0.0	0.0	0.0	Luethy, Alexandrov, Bass & Solovyev

it was decided to require more than two predictions to qualify for the overall ranking discussed in the next section.

Picking Winners?

Tables IX and X sort the group averages by the ACrct score. While this gives a ranking that fits well with those chosen to speak at the meeting, it is too one-dimensional to allow a full appreciation of sustained performance.

To pick winners who did well in both alignment accuracy and fold recognition, I plot the average ACrct against the average TSpc for both the raw and normalized data (Figs. 3a,b). The five groups chosen to write articles on threading (Murzin, Sippl, di Francesco, Eisenberg, and Karplus) stand out clearly on the plot, which does not allow for target difficulty (Fig. 3a). Two other groups that do almost as well are Alexandrov and Luethy. The fact that Alexandrov himself is a member of both groups complicates the

TABLE VIII. Better Results for Target T0038 Sorted by ACrct

Entry Code	Conf	TSpc	ACrct	ARms	Shft	Covr	Predictors
T0038FR585	90	75.0	22.5	8.0	2.0	85.3	Murzin & Bateman
T0038FR610	85	19.3	18.5	14.1	8.8	67.5	Eisenberg, Weiss, Rice & Fischer
T0038FR612	100	40.7	17.9	9.7	39.5	71.2	●● Bryant, Hogue, Madej, Marchler-Bauer & Ohkawa
T0038FR575	50	32.7	12.8	11.4	17.8	67.3	Jones
T0038FR614	70	25.0	6.0	19.1	75.1	38.1	Abagyan & Batalov
T0038FR513	100	15.3	5.1	15.4	8.1	29.6	Sternberg, Bates, Russell, Saqi & Sayle
T0038FR595	100	28.5	5.0	16.1	6.7	28.1	Lengauer, Mevissen, Thiele & Zimmer
T0038FR609	90	28.5	1.4	15.7	7.7	28.5	Alexandrov & Zimmer
T0038FR616	33	0.0	1.0	15.4	16.4	79.3	Lathrop, Rogers, Smith & White

TABLE IX. Mean[1] Results for Groups[2] Sorted by ACrct

Rank	Nent	Conf	TSpc	ACrct	ARms	Shft	Covr	Predictors
1	2.0	100	64.2	59.8	13.1	1.8	54.7	●● Dixon & Thomas
2	7.0	100	36.9	30.2	7.0	3.0	34.0	●● Sippl, Floeckner, Domingues & Jaritz
3	4.9	81	60.4	25.5	5.1	1.7	56.5	●● Murzin & Bateman
4	6.4	71	40.9	20.8	9.6	3.0	37.0	Alexandrov & Zimmer
5	7.5	57	42.7	20.0	10.0	13.6	42.1	●● Eisenberg, Weiss, Rice & Fischer
6	4.0	100	57.5	19.4	10.8	5.5	45.8	●● Di Francesco, Geetha, Garnier & Munson
7	7.2	90	31.7	19.0	8.9	1.1	26.7	●● Karplus, Sjölander, Barret, Cline, Haussler, Hughey, Holm & Sander
8	8.8	67	33.1	17.1	12.1	3.9	31.4	Luethy, Alexandrov, Bass & Solovyev
9	9.3	77	33.9	14.4	12.4	7.5	29.9	Hubbard, Park & Reinhardt
10	9.0	75	26.2	13.7	11.2	12.4	28.7	Lengauer, Mevissen, Thiele & Zimmer
11	6.5	50	22.1	13.5	13.2	30.8	37.5	Jones
12	6.0	100	28.2	12.2	14.6	31.4	28.0	Taylor & Munro
13	8.7	86	30.0	10.7	12.2	2.8	22.7	Valencia, Pazos, Olmea & Rost
14	11.4	87	19.8	9.6	16.0	7.1	23.5	Sternberg, Bates, Russell, Saqi & Sayle
15	9.5	73	15.9	9.1	12.8	2.6	21.8	Solovyev
16	6.4	58	34.9	8.9	13.4	20.9	43.3	Bryant, Hogue, Madej, Marchler-Bauer & Ohkawa
17	3.4	68	20.6	6.2	9.0	0.3	18.9	Kim & Dubchak
18	11.0	28	12.8	5.5	15.8	21.9	22.2	Coulson
19	0.8	40	19.7	4.2	14.0	2.7	15.4	Abagyan & Zhou
20	6.0	100	6.1	3.5	13.5	30.7	26.3	Honig, Yang & Xiao
21	1.6	80	10.9	2.6	19.1	75.1	16.7	Abagyan & Batalov
22	12.0	100	10.2	2.5	19.3	28.2	21.8	Torda, Huber, Dyer & Lu
23	2.2	73	17.0	1.6	22.7	16.6	11.6	Moult, Milash, Braxenthaler, Pedersen & Samudrala
24	10.0	71	2.3	1.6	18.1	19.8	11.6	Elofsson
25	8.0	44	6.4	1.2	12.8	14.0	25.7	Lathrop, Rogers, Smith & White

[1]The mean value of each index was obtained by averaging over all the predictions for a given group of predictors. These groups are as defined from the prediction entries; no attempt has been made to merge predictions that come from the same laboratory with different authors. In doing this averaging, care was taken to weight the quantities by the confidence the predictors placed on their prediction (the percentage Conf value divided by 100). The total Conf/100 value is the effective number of entries, Nent, for the particular group.
[2]Here and in Table X, results are only presented for the 25 groups that do best overall. In Figures 3, 4, and 5, we only present results for the better scoring groups.

assessment in a way that has not been properly allowed for here: it might be argued that one person should not be able to be part of two independent groups. It turns out that the predictions of the Alexandrov and Zimmer group were in fact the automatic predictions of the 123D program,[10] whereas Alexandrov is actually a member of the Luethy group and Zimmer is a member of the Lengauer group (Zimmer, personal communication).

Use of the average ACrct and TSpc values normalized for target difficulty produces a different picture (Fig. 3b) in which four groups stand out (Murzin, Eisenberg, Hubbard, and Bryant) followed by Jones and Sippl. Combining these results with those without normalizing for target difficulty (Fig. 3a) gives the following overall ranking.

The clear winner is Murzin and Bateman. This group did the best on three out of the seven targets

TABLE X. Mean Results for Groups Normalized and Sorted by ACrct

Rank	Nent	Conf	TSpc	ACrct	ARms	Shft	Covr	Predictors
1	4.9	14	4.36	4.93	0.26	0.29	2.89	●● Murzin & Bateman
2	7.5	5	1.34	2.13	0.41	0.66	1.21	●● Eisenberg, Weiss, Rice & Fischer
3	9.3	7	0.90	2.05	0.51	0.78	0.74	Hubbard, Park & Reinhardt
4	6.4	6	1.16	1.87	0.47	0.77	1.23	Bryant, Hogue, Madej, Marchler-Bauer & Ohkawa
5	2.0	51	1.67	1.46	0.53	0.52	1.28	●● Dixon & Thomas
6	6.5	4	0.62	1.11	0.49	1.15	1.66	Jones
7	7.0	15	0.70	0.98	0.40	0.34	0.59	●● Sippl, Floeckner, Domingues & Jaritz
8	4.0	26	1.10	0.80	0.62	0.98	0.83	●● Di Francesco, Geetha, Garnier & Munson
9	1.6	41	0.58	0.75	0.78	5.00	0.61	Abagyan & Batalov
10	6.4	8	0.91	0.73	0.43	0.32	0.71	Alexandrov & Zimmer
11	8.8	6	1.07	0.64	0.50	0.42	0.89	Luethy, Alexandrov, Bass & Solovyev
12	14.0	8	1.39	0.58	0.56	1.75	1.12	Rost
13	7.2	12	0.72	0.58	0.41	0.38	0.52	●● Karplus, Sjölander, Barret, Cline, Haussler, Hughey, Holm & Sander
14	9.0	7	0.63	0.54	0.47	0.81	0.69	Lengauer, Mevissen, Thiele & Zimmer
15	11.4	7	1.63	0.45	0.54	0.80	1.23	Sternberg, Bates, Russell, Saqi & Sayle
16	6.0	18	0.68	0.38	0.61	2.10	0.56	Taylor & Munro
17	3.4	14	0.49	0.36	0.48	0.22	0.37	Kim & Dubchak
18	9.5	6	0.74	0.35	0.53	0.80	0.66	Solovyev
19	8.7	9	0.70	0.34	0.52	0.38	0.47	Valencia, Pazos, Olmea & Rost
20	6.0	18	0.23	0.27	0.47	1.50	0.75	Honig, Yang & Xiao
21	3.0	34	0.93	0.21	0.60	1.95	1.11	Sanejouand
22	11.0	1	1.10	0.18	0.52	1.30	1.24	Coulson
23	8.0	3	0.37	0.18	0.61	0.51	1.31	Lathrop, Rogers, Smith & White
24	11.4	7	0.38	0.16	0.54	2.04	0.92	Barton & Copley
25	0.8	20	0.51	0.13	0.80	0.81	0.42	Abagyan & Zhou

(T0002, T0014, and T0038) and their outstanding performance is also evident in Figures 3a and b. In the second tier there are two groups: Sippl, Floeckner, Domingues, and Jaritz and Eisenberg, Weiss, Rice, and Fischer. The Sippl group stands out when target difficulty is not taken into account (Fig. 3a) and the Eisenberg group stands out when it is taken into account (Fig. 3b). However, both do well in either situation. The Sippl group did best for T0031, an easy target, whereas the Eisenberg group did relatively well for T0020 and T0038, both difficult targets.

In the third tier there are six groups who do very well in either the un-normalized (Fig. 3a) or normalized (Fig. 3b) analysis and includes di Francesco, Karplus, Alexandrov, Bryant, Hubbard, and Jones. The selection of two of these six groups to contribute articles was particularly difficult, even after eliminating the Bryant group, who did the threading evaluation. The di Francesco group was selected because they did consistently well in fold recognition with a high threading specificity score. The Karplus group was selected because they did well in model accuracy with good overall ARms scores (Figs. 4a,b). It must be emphasized that there is a continuum of scores and other groups, such as Luethy, Lengauer, and Taylor, also do very well, making these decisions marginal.

How much do these overall results depend on the choice of the primary evaluator indices TSpc and ACrct? To test this, I plot the average modeling accuracy, ARms, against the average TSpc, ignoring and allowing for target difficulty (Figs. 4a,b). Although both ARms and ACrct depend on how well the predicted alignment matches the target structure to the database folds, ARms does not depend on the structural alignment produced by Dali or VAST, whereas ACrct does. Figure 4a shows that when target difficulty is not taken into account, the top six groups selected by ARms and TSpc are just the same as found for the plot of ACrct against TSpc (Fig 3a).

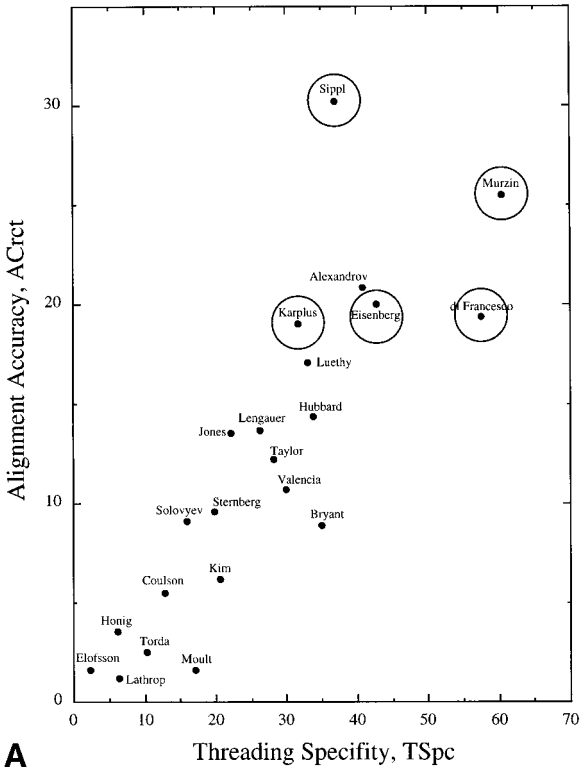
Comparing Figures 4b and 4a shows that when the scores for each target are normalized the effect on relative rank is smaller than that seen in the comparison of Figures 3a and 3b. The five top ranking groups are Murzin, Eisenberg, Sippl, Karplus, and Alexandrov. Only di Francesco has moved to a much worse position. It is interesting that the outstanding performance of Bryant and Hubbard seen in Figure 3b is not apparent in Figure 4b.

DISCUSSION

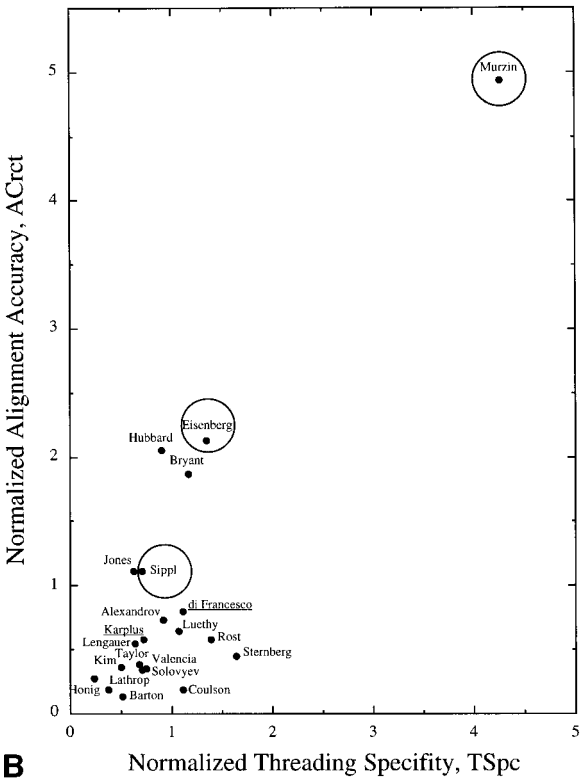
Assessing Sustained Performance

The most difficult task facing the assessor is to choose the overall winners. In this assessment, I calculated the sustained performance as a weighted

average of the performance on each target predicted. This scheme has a limitation in that it rewards those predictors who did very well on the easier targets.



A



B

Even when the targets are weighted for difficulty, the easy targets still count just as much as the hard targets. Were the same scheme applied to other competitions, such as the Olympic Games, it would be equivalent to averaging the times from different events. Sustained performance at the Olympic Games is measured by counting the number of medals (perhaps weighting for gold, silver, and bronze), or, for compound events like the pentathlon, is measured by adding the scores for the individual events. Counting "medals" for the seven different threading targets ranked by the alignment accuracy (somewhat arbitrarily assigning 3 points to a first place, 2 to a second place, and 1 to a third place), ranks the eight top groups as follows (total score in parenthesis after each name): Murzin (9), Eisenberg (5), Jones (4), Bryant (4), Sippl (3), Alexandrov (3), and Rost (3). While Murzin is still clearly top, Jones is now highly ranked. Ranking winners on individual targets by other criteria may affect these scores, as would giving more weight to the more difficult targets.

Another possible measure of sustained performance is to simply give a score of 1 for appearing in the tables of the better results for each of the seven targets (Tables II–VIII). By this measure, the ranking is as follows (number of targets for which the group did well in parenthesis): Jones (7), Eisenberg (6), Murzin (4), Bryant (4), and Sternberg (4). By this measure, Jones emerges as the winner, with remarkably consistent performance for all the seven targets!

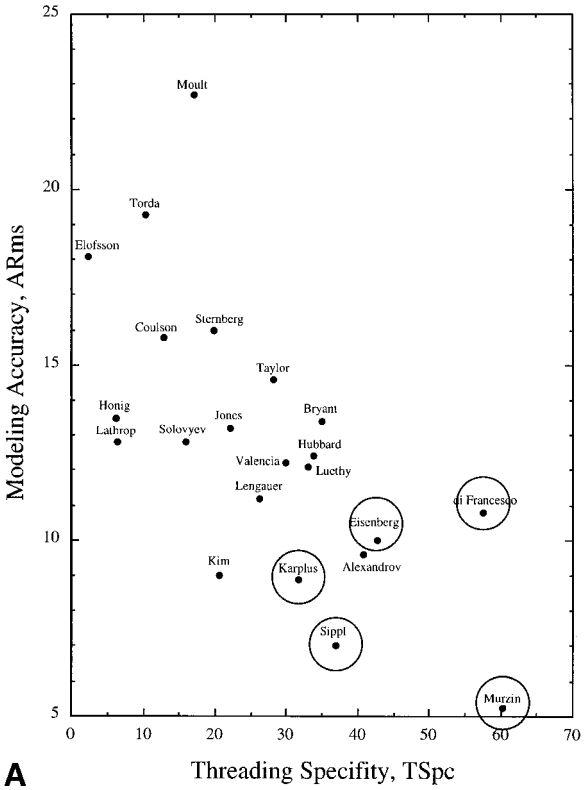
It is tempting to contend that there is no fair way to measure sustained performance, as one is really adding apples and oranges. This argues strongly for an independent assessor. In fact, one can go further and argue that it requires that the assessor be eliminated and assessment automated, as evaluation is at present. This requires one or more precisely defined criteria for evaluation of sustained performance. This is a problem that is addressed in the following paper, involving a collaboration of both the threading evaluators and the assessor.[11]

Progress With Threading Predictions

It is clear to this assessor that progress on threading prediction has been very significant in the two years since the first CASP meeting. This is evidenced

Fig. 3. (**A**) The overall average TSpc and ACrct scores of the high-ranking groups, with each point being labeled by the name of the group (data from Table IX). Seven groups stand out from the others in terms of their ACrct and TSpc scores—Sippl, Murzin, di Francesco, Eisenberg, Alexandrov, Karplus, and Luethy. In Figures 3 and 4, the circles mark the groups chosen to write articles on their work. (**B**) The overall average TSpc and ACrct scores after normalizing for target difficulty by scaling so that the mean normalized TSpc and ACrct scores of each target is 1.0 (data from Table X). Four groups stand out from the others in terms of their normalized ACrct and TSpc scores—Murzin, Eisenberg, Hubbard, and Bryant.

by the very large number of groups that succeed in recognizing the easier targets. It is clear that a wide variety of different threading methods used by many different groups do much better than simple sequence alignment against the database of known folds. As this is an assessment of performance rather than a review of threading methods, the reader interested in methodological details is strongly urged to consult the individual articles in this issue of *Proteins* and also to the publications of the predictors mentioned here. The results of Murzin and Bateman are particularly noteworthy, as their predictions were done manually, relying on Murzin's unparalleled knowledge of protein structure and the associated literature. The superior performance of this method compared to all the other computerized methods suggests that by incorporating more knowledge automatic methods should be able to improve.



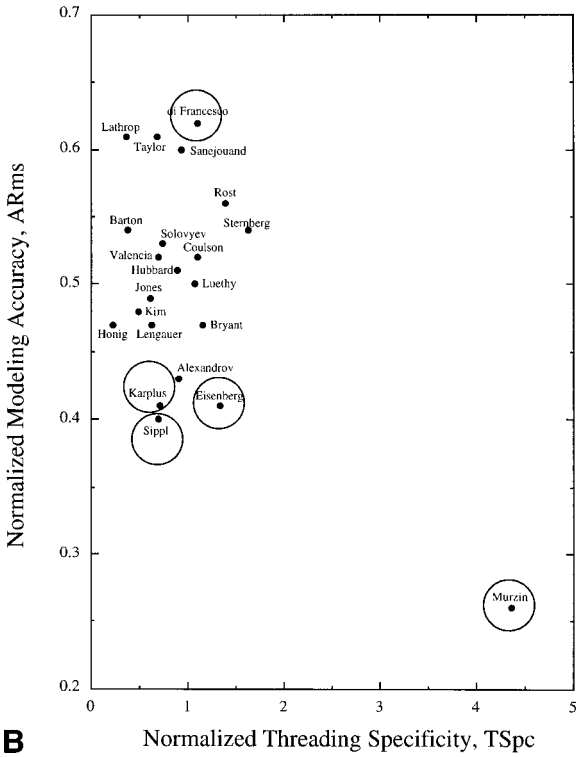


Fig. 4. (**A**) The overall average TSpc and average ARms scores of the high-ranking groups, with each point being labeled by the name of the group (data from Table IX). Six groups stand out from the others in terms of a low ARms and high TSpc score—Murzin, Sippl, Karplus, di Francesco, Eisenberg, and Alexandrov. (**B**) The overall average TSpc and average ARms scores after normalizing for target difficulty (data from Table X). One group is clearly separated from all others (Murzin) and there are another four groups that are separated from the rest—Eisenberg, Sippl, Karplus, and Alexandrov.

The best predictions were really quite remarkable, both in terms of the alignment accuracy (ACrct) and modeling accuracy (ARms). Models with ARms less than 6 Å were predicted for T0002, T0004, T0014, and T0031. In fact, the average ARms for all the predictions of the Murzin group was 5.1 Å! The alignments of these good models were also correspondingly good with high alignment accuracy (ACrct) and small alignment shift (Shft). The best individual models predicted for three targets were exceptionally good, with ARms values of 2.83 Å, 2.97 Å, and 4.16 Å and matching lengths of 64, 65, and 202 residues for targets T0002, T0004, and T0031, respectively (predicted by the Murzin, Jones, and Sippl groups, respectively).

A more qualitative comparison with the threading assessment done for CASP1[12] is difficult, as different evaluation indices are used this time. Nevertheless, some comparisons can be made. There were many more teams of predictors this time (over 30 vs. 9) but the number of targets is smaller (15 vs. 21) and fewer of these targets have recognizable folds (7 vs. 11). The fold-recognition results are much better this time, with one group (Jones) making the Better Results tables for all seven targets, whereas before the best score was 5 out of 11 (also Jones). Comparison of the alignment and modeling accuracy is more difficult, as the criteria used for CASP1 were much less quantitative. Here, the best models for the three easy targets had RMS deviations of between 2.9 Å and 4.2 Å. For the two easiest targets (T0004 and T0031), the predicted alignment matched the actual structural alignment perfectly over more than 80% of the matching residues! The choice of contributors

for CASP1 is also interesting (Sippl, Bryant, Jones, Nishikawa, and Hubbard), including as it does four groups who also did well this time (Sippl, Bryant, Jones, and Hubbard). In CASP1, this choice was primarily by number of predictions rather than by prediction quality (Wodak, personal communication).

Lessons for the Future

The target sequences need to be checked more carefully to ensure that they are indeed valid threading targets and have no significant sequence similarity to any database protein. If only part of the sequence is a valid threading target, this needs to be stated clearly. In CASP2, there were two homology modeling targets, T0008 and the first domain of T0002, that were really homology modeling targets.

Thought needs to be given to the participation of an investigator in different groups, as it should not be possible for a predictor to do better by simply submitting more predictions in different collaboration!

All technical issues associated with the hiding of predictor names, duplicate entries, NONE predictions, and proper averaging should be done automatically by the evaluators. This would allow the assessor to concentrate more on judging the entries, both in terms of the individual predictions as well as the sustained performance.

Finally, one needs to find a way to reward those groups who correctly conclude that a target does not match any fold in the database. This would mean that each predictor should be encouraged to submit an entry for those targets for which fold-recognition fails placing high confidence in the "NONE" fold. This issue is addressed in the paper by Marchler-Bauer et al.[11]

CONCLUSION

The process of this assessment has lead to a number of conclusions that I was not sufficiently aware of when agreeing to assess the CASP2 threading predictions. First, while it is possible to say who did best on a particular target, any attempt to decide overall winners is fraught with very real difficulties. It is now clear that early attempts made to rank groups based on their sustained performance were naive. This realization, which may have been obvious to wiser scientists, was not clear until the final stages of the analysis and after I had selected both the speakers and the groups asked to contribute articles. While those chosen did do well, it is clear that the choice is not robust: with slightly different criteria the list of overall winners would change.

Second, in spite of all claims to the contrary, we are all very subjective in judging how well we are doing. This observation applies equally to the assessor trying to justify his decisions, the evaluators trying to prove the value of their criteria, and the predictors trying to show that their predictions are correct. If there were any doubts concerning the need for critical assessment of structure prediction, they were completely dispersed by my experiences. The CASP process is essential for this field: without it, progress on protein structure prediction will be very severely hampered.

ACKNOWLEDGMENTS

I would like to acknowledge the moral support and encouragement given me by many colleagues at the meeting, which was a time of considerable stress.

REFERENCES

1. Marchler-Bauer, A., Bryant, S.H. Measures of threading specificity and accuracy. Proteins Suppl. 1:74–82, 1997.
2. Holm, L., Sander, C. Protein structure comparison by alignment of distance matrices. J. Mol. Biol. 233:123–128, 1993.
3. Holm, L., Sander, C. Alignment of 3-dimensional protein structures: Network server for database searching. Methods Enzymol. 266:653–662, 1996
4. Taylor, W.R., Orengo, C.A. Protein structure alignment. J. Mol. Biol. 208:1–22, 1989.
5. Orengo, C.A., Taylor, W.R. SSAP: Sequential structure alignment program for protein–structure comparison. Methods Enzymol 266:617–635, 1996.
6. Gibrat, J.-F., Madej, Y., Bryant, S.H. Surprising similarities in structure comparison. Curr. Opin. Struct. Biol. 6:377–385, 1996.
7. Feng, Z.K., Sippl, M.J. Optimum superimposition of protein structures: Ambiguities and implications. Folding and Design 1:123–132, 1996.
8. Godzik, A. The structural alignment between two proteins: Is there a unique answer. Protein Sci. 5:1325–1338, 1996.
9. Orengo, C.A., Brown, N.P., Taylor, W.R. Fast structure alignment for protein databank searching. Proteins 14:139–167, 1992.
10. The 123D program at http://cartan.gmd.de/ToPlign.html or http://www-lmmb.ncincrf.gov:80/~nicka/runl23D.html
11. Marchler-Bauer, A., Levitt, M., Bryant, S.H. A retrospective analysis of CASP2 threading predictions. Proteins Suppl. 1:83–91, 1997.
12. Lemer, M.-R., Rooman, M.J., Wodak, S.J. Protein structure prediction by threading methods: Evaluation of current technologies. Proteins 23:337–355, 1995.