PROTEINS: Structure, Function, and Genetics Suppl 3:231–237 (1999)

Some Measures of Comparative Performance in the Three CASPs

Česlovas Venclovas,[1] Adam Zemla,[1] Krzysztof Fidelis,[1] and John Moult[2]*
[1]*Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California*
[2]*Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, Maryland*

ABSTRACT **Performance in the three Critical Assessment of protein Structure Prediction (CASP) experiments has been compared in the areas of alignment accuracy for models based on homology and three-dimensional accuracy for models produced by using ab initio prediction methods. The homologous models span the comparative modeling and fold-recognition regimes. Each CASP target is assigned a relative difficulty based on the extent of sequence identity and the degree of structural overlap with the best available template. There is a clear improvement in alignment accuracy between CASP1 and CASPs 2 and 3 over much of the difficulty scale but no apparent improvement between CASP2 and CASP3. Encouragingly, the best ab initio models of small targets are clearly more accurate in CASP3 than in CASPs 1 and 2. Proteins Suppl 1999;3:231–237.** **Published 1999 Wiley-Liss, Inc.†**

Key words: protein structure prediction; community wide experiment; CASP

INTRODUCTION

Now that three Critical Assessment of protein Structure Prediction (CASP) experiments have been completed, it should be possible to determine whether progress is being made in our ability to produce accurate and useful models of protein structure. This study describes one of the three attempts to begin to do that reported in this issue.[1,2]

There are three principal questions to be addressed in devising numerical measures of progress:

Difficulty Scale: All protein structures are not equally difficult to model. At one end of the spectrum, CASP data show that a backbone copied from a template with high sequence identity (> 60%) to the target protein in almost all cases gives an RMS (root mean square) error on Cα atoms of less than 1Å. At the other end, it is still almost impossible to produce useful models of proteins with no detectable fold relationship to a known structure. In between the two extremes, there is a gradual decrease in model quality as a function of decreasing sequence and structural similarity between target and template. Thus, to compare performance in different CASPs, it is necessary to devise some scale of relative difficulty for all the targets. Below, we describe one such scale. More work is needed in this area.

Choice of Evaluation Criteria: Many different numerical measures have been introduced over the three CASPs; which are most useful for measuring progress? We use a subset of the methods developed at the Livermore Prediction Center. These methods are described elsewhere in the issue,[3] and further details are available at the CASP web site.[4] For this first attempt at measuring progress, we have focused on two aspects of performance — the quality of alignments and the quality of the three-dimensional models produced by "ab initio" prediction methods.

Choice of Models To Be Included: For each target in CASP, many models are typically submitted. In assessing progress, one must decide which of these to consider: Is the performance of the very best people in the field most important, or should one measure the general state of the art? Is it important which model the predictor felt was his or her best for a particular target, or should one consider the model that actually turned out to be the best, irrespective of how the predictor rated it? Should one look at performance for predictors across all targets, or consider the best result on a particular target, irrespective of which team submitted it? We have compiled data for three of these possibilities: First, including only the models for which a predictor expressed the most confidence. In CASP3, this is the so-called "Model 1"s, in CASP2, the model given the highest weight, and in CASP1, the first model in a submitted list. We refer to this as the "Model 1" progress evaluation. Second, taking the best model submitted by any predictor for a given target, independent of the confidence the predictor expressed in it. Third, calculating an average score over the best models submitted for up to the six best performing groups on each target (for CASP1, there were not always predictions from six groups available).

DIFFICULTY SCALE

For prediction targets closely related to a known structure, the percentage of sequence identity is a reasonable indication of difficulty in producing a model. However, for distant relationships and nonhomologous similar folds, this signal becomes much less useful. Instead, the fraction of the target structure that can be superimposed on an available template is a more meaningful measure of

Grant sponsor: DOE; Grant number: DE-FG02-96ER62271.
Dr. Venclovas's permanent address is Institute of Biotechnology, Graičiūno 8, 2028 Vilnius, Lithuania.
*Correspondence to: John Moult, Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, MD 20850. E-mail: jmoult@carb.nist.gov
Received 10 June 1999; Accepted 15 June 1999

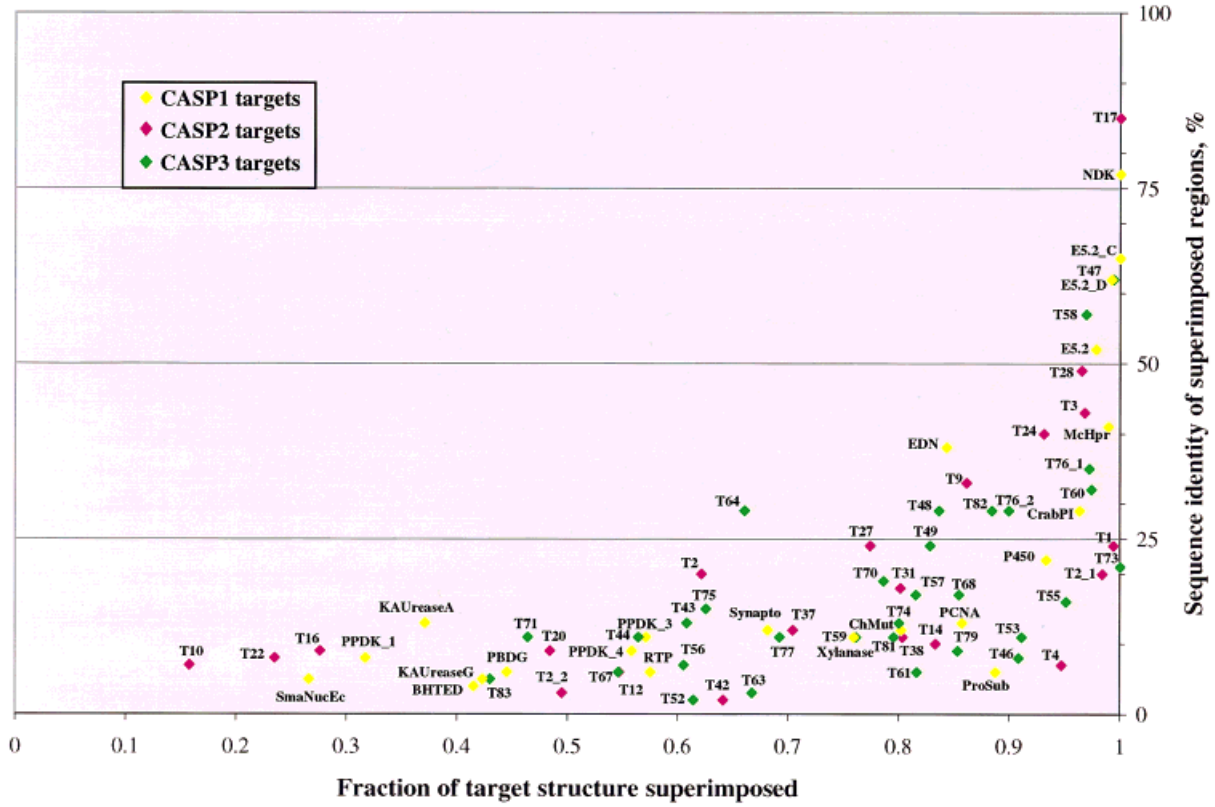


Fig. 1. Relative difficulty of targets in the three CASP experiments, displayed as a function of the percentage sequence identity between the target and the best available template (vertical axis) and the fraction of the target structure that can be superimposed on the template (horizontal axis). Comparative modeling targets tend to cluster at the right-hand side of the plot, with easiest (high sequence identity) at the top. Fold-recognition targets are spread out from right to left, with the most difficult (usually analogous relationships) at the far left. Targets from each CASP span the full range of difficulty.

difficulty. Work by Marchler-Bauer and Bryant has shown that in CASP2[5] and in CASP3,[1] for fold-recognition targets, model quality varies as combination of these two factors. We have used a similar metric of difficulty, extended to cover the comparative modeling and the fold-recognition areas of CASP. (Targets that are considered to be tests of progress for ab initio methods are treated separately; see below.)

The fraction of residues that can be superimposed between a target and the best available template is based on a sequence-independent structure superposition performed with DALI.[6] DALI is one of three structure comparison methods used in CASPs 2 and 3, providing continuity. In CASP3, PROSUP[7] results were used in evaluation tables distributed at the meeting, and so were given a high weight. Tests with the CASP3 targets showed that the ranking obtained with DALI is very similar to that suggested by PROSUP superpositions, except that PROSUP was slightly superior in finding remote templates. (For example, it identified structural similarity for T0054 and T0085).[8] For CASP2 and CASP3, possible templates were taken from the data compiled by Liisa Holm and available at the CASP web site.[4] For CASP1, the best templates available at that time were identified as follows: Each target structure was compared with the library of representative structures now available on the DALI server.[9] The set of structural relationships so detected was expanded by performing a sequence search with SSEARCH[10] of each template against all PDB structures available at the time of CASP1. Each significant sequence relative (E value $< 10^{-5}$) was then superimposed on the target to find the best template.

For each target, only templates that DALI superimposed with less than 3.5Å RMS deviation were considered, and the one that produced the largest number of structurally equivalent residue pairs were selected as the best. Sequence identity for the superimposed region was taken directly from the DALI output.

Figure 1 shows the distribution of the fraction of superimposable residues and percentage sequence identity for all targets from CASPs 1, 2, and 3 where DALI found a template. Seventeen ab initio or very remote fold-recognition targets are not included. Targets in the comparative modeling regime tend to cluster at the right side of the plot, with the "easiest" (high sequence identity) targets at top right. Targets that have low sequence identity are well spread on the horizontal axis, reflecting different levels of superimposability on the best available template. The most difficult targets are at the lower left corner. Inspection of this distribution shows that, in general, targets

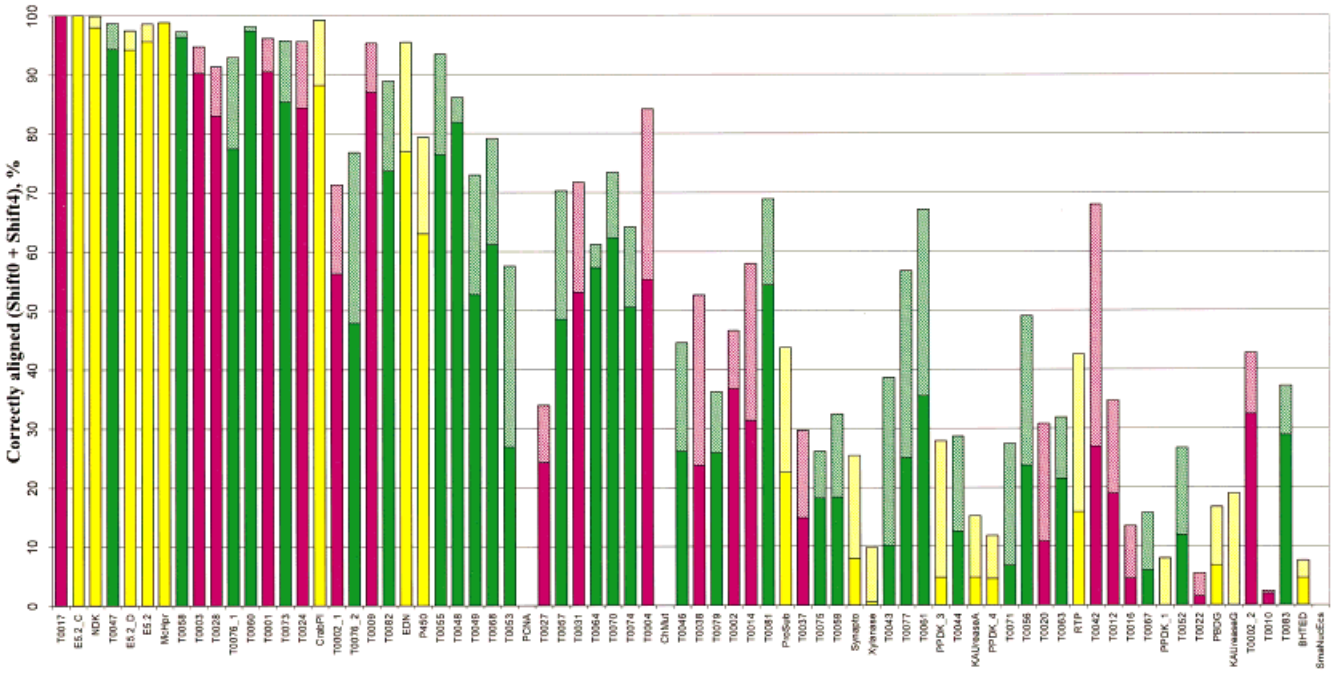


Fig. 2. Fraction of residues correctly aligned between the target structure and the best model for each target. Yellow, CASP1; purple, CASP2; green, CASP3. Full bars represent the fraction of correctly aligned residues, and hatched bars represent the additional fraction of residues in error by not more than four residue positions. The targets are arranged left to right, starting at the least difficult. Alignment accuracy falls steadily with increasing difficulty of targets. For the more difficult targets, it is clear that CASPs 2 and 3 performance is superior to CASP1, but there is no easily discernible difference between 2 and 3.

have relative positions approximately where CASP experience suggests they should be. In particular, targets classified as ab initio, where included, are at lower left, and fold-recognition targets span the lower part of the distribution, with "fold" relationships tending to be on the left, and "superfamily" relationships toward the right. There are a number of exceptions to the intuitive order, but we consider that on average this measure is meaningful enough to be useful. Because the targets do not fall on any single line, a two-dimensional representation of difficulty should ideally be used. As a reasonable first approximation, we have used a one-dimensional ranking of difficulty. That is, the difficulty of each target is expressed as a linear combination of ranks by structure superimposability and sequence identity (RANK_ STR_ALN + RANK_ SEQ_ID)/2, where RANK_STR_ALN is the rank of the target along the horizontal axis, and RANK_SEQ_ID is the rank along the vertical axis.

ALIGNMENT ACCURACY

A major determinant of model quality in both the comparative modeling and fold-recognition regimes is the effect of errors in mapping the target sequence onto template structures. We measure such alignment quality from the sequence-independent superposition of the model and target structures generated by DALI. Alignment accuracy was calculated from these superpositions by considering each position in the sequence in turn, and asking whether the Cα atoms of equivalent residues in the two structures are within 3.8Å, and also checking that neither atom has a closer Cα. Residues passing these two tests are counted as correctly aligned. Note that this is a more stringent definition than used by the Sippl group, where the cutoff distance is 5Å, and there is no requirement that the equivalent residues be the closest neighbors.[11] The Sippl procedure also searches for alternative alignments around the initial superposition. As a consequence of these factors, the number of residues considered correctly aligned is smaller for our procedure. Also note that a direct comparison of a model structure with the experimental target structure is different from the procedure most used in CASP2, where alignment accuracy was measured from the relationship between the model and template structures.

Figure 2 shows the results for the "best" model for each target. Consider first the left-hand part of the plot, up to and including target 68. All the targets here are in the comparative modeling regime, ranging from 85% identity to 17% identity, and the difficulty order is dominated by the sequence identity term. The very high identity models have essentially zero alignment errors, but errors increase rapidly as the degree of sequence identity declines, down to typically only 70% correct for the more difficult cases. There is no apparent difference between the three CASPs.

The next block of predictions, spanning T0053 to T0004, is a mixture of low identity comparative modeling targets and "easy" fold-recognition targets. Alignment accuracies are around 60%. There were no CASP1 targets displayed in this interval, and the CASP2 and CASP3 results appear similar.

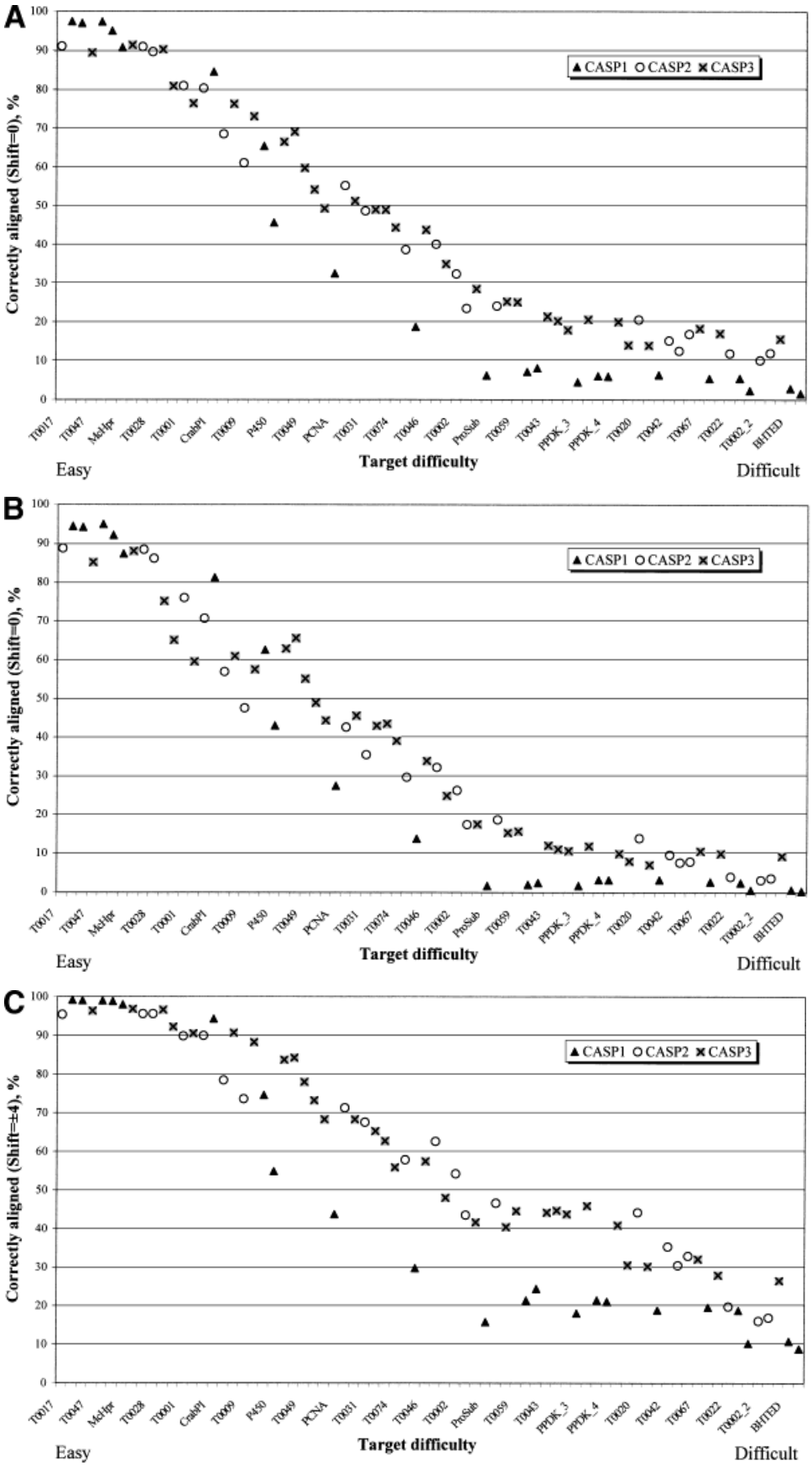


Fig. 3. Fraction of residues correctly aligned between the target structure and the best model for each target (**A**) and the average over the six best models (**B**). To make the trends more visible, each data point is smoothed by averaging over itself and the two neighboring points on each side. Both plots show that performance in CASPs 2 and 3 are approximately equal, and significantly better than in CASP1. Averaging over the six best models gives lower scores, indicating there are only a few outstanding predictors. **C,D:** The fraction of residues aligned to within ± four residues, C for best models and D for averages over the best six. Scores are substantially higher, allowing this margin of error, but the trends over the CASPs are very similar.

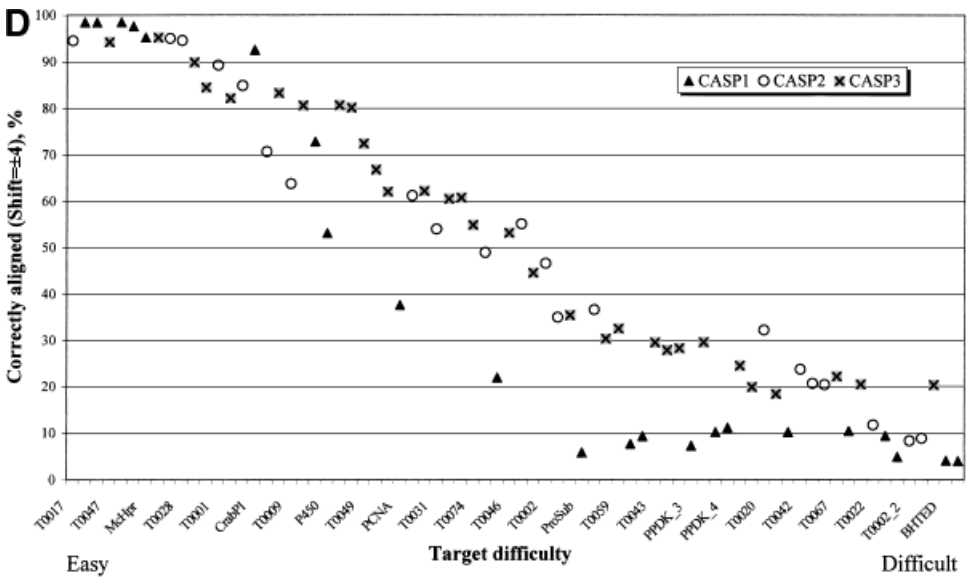


Figure 3. (Continued.)

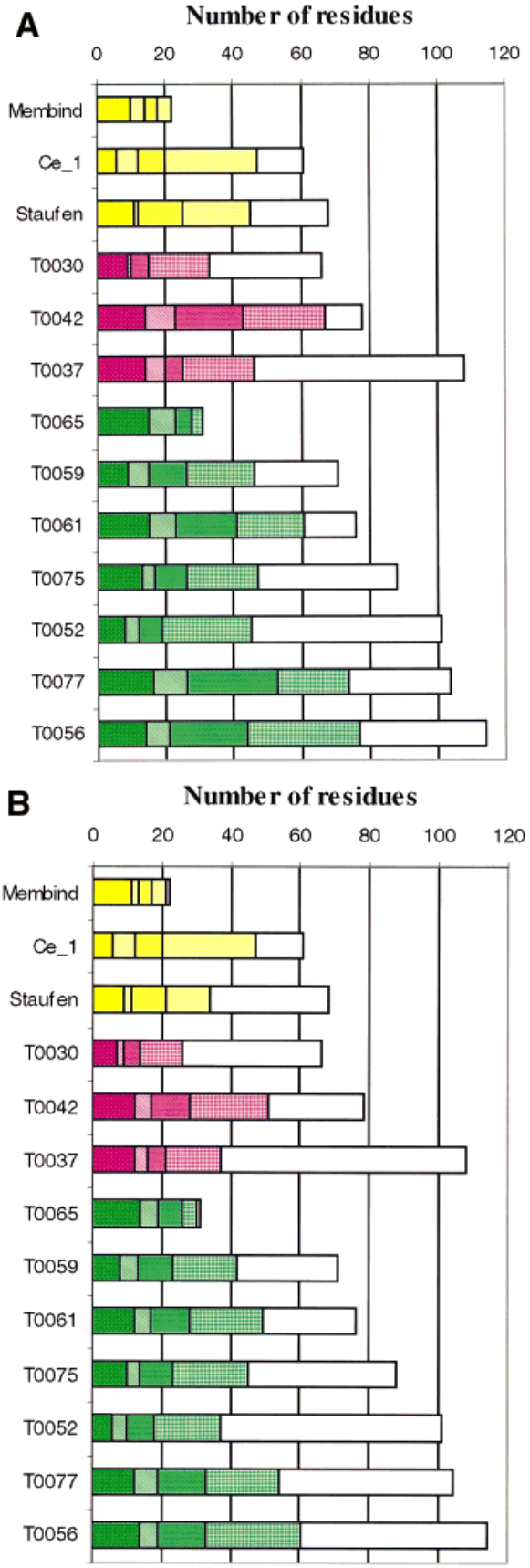
The rest of the range covers increasingly difficult fold recognition targets, and there is a corresponding gradual decrease in alignment accuracy. It is clear that CASP1 accuracy is significantly worse than CASPs 2 and 3. Given the general noisy nature of the distribution, it is not possible to see any significant difference between CASPs 2 and 3 by this measure. There are some interesting features in this region though: The CASP2 target T0042, which stands out as a high-quality result for its difficulty rating, was treated as an ab initio target by all but one of the successful predictors. T0083 is a case in which one domain has a superfamily relationship to a known structure, but the rest of the molecule is a new fold. Therefore, it ranks as quite difficult on our scale, but many predictors produced good models for the superfamily related region, hence, the relatively high alignment scores.

Displaying the raw data allows one to appreciate the details of the distribution but not to easily see overall trends that may be significant for detecting progress. Figure 3 shows the same data, but smoothed, i.e., the score for each target is the average over itself and the closest two targets from the same CASP on each side. Figure 3A shows the same data as Figure 2. It is immediately clear that starting in the moderately difficult comparative modeling region (around P450) and extending through to the most difficult targets, the CASP1 results are significantly worse than those of the other CASPs, but that CASP2 and 3 are essentially the same. Figure 3B shows the results allowing up to four-residues alignment error. For a large section of the difficulty range, scores are substantially better, but the overall relationship between the three CASPs is the same as in Figure 3A. Figure 3C,D again shows the fraction correctly aligned and aligned within ± four residues, respectively, but averaged over the six best predictions on each target. Scores are generally lower in these plots, indicating that there were a few outstanding predictors, but again CASP1 is the worst, and CASP2 and 3 are very similar. The "Model 1" plots lead to the same conclusions and, therefore, are not shown here.

AB INITIO PREDICTION PERFORMANCE

We have compared ab initio prediction quality over the three CASPs. The comparison is restricted to the relatively small targets (less than approximately 120 residues) that were in the "difficult" or "impossible" categories for fold recognition and considering only predictions believed to be made by using ab initio methods. These small targets are the ones for which numerically intensive ab initio methods can be applied. Because ab initio prediction methods are not yet powerful enough to produce good models of full-size proteins, it is necessary to use evaluation procedures that identify any accurately predicted substructures. These substructures may not necessarily be composed of contiguous regions of the sequence. We have used the Global Distance Test (GDT), described in the Livermore Methods article in this issue.[3] The algorithm finds the maximum number of residues for which the distance between the target and corresponding model Cα is less than some threshold, in a sequence-dependent superposition. In looking at the results, it should be borne in mind that a distance threshold is a stricter criterion than an RMS deviation, in the sense that the RMS deviation of a set of residues is usually substantially less than the distance threshold used to define the set. (See the GDT data on the CASP web site[4] for examples of the relationship between a distance threshold and RMS deviation; factors of up to two are not uncommon.) We consider distance thresholds of 1, 2, 4, and 8Å.

Figure 4A shows the results for the best model submitted for a given target, and Figure 4B shows an average over up to six best predictions from different groups. "Model 1" results are slightly worse than "best model," but in this most difficult category, it is reasonable to focus on the best. In the "best model" plot, there is only one prediction in CASPs 1 and 2 with more than 40 residues below the 4Å distance threshold, and the others all have less than 25. The high scoring one is the CASP2 target, T0042, an all α protein. This target was recognized at the



time as the single case for which meaningful ab initio predictions were made (there was also one high-quality threading prediction of the same target). In CASP3, three targets have 40 residues or more below this threshold and also have over 60 residues below 8Å. The best performance is for the CASP3 target T0077, a mixed α/β structure, with more than 50 residues below 4Å. In the plot showing the average over the six best predictions, two of the more successful CASP3 targets have significantly lower scores, indicating that only a few predictors could produce the best results. The distinct improvement seen in going from CASPs 1 and 2 to CASP3 is quite encouraging. However, all but the smallest targets still have large "white" bars. That is, regions that are poorly modeled, so, not surprisingly, there is still a long way to go. The two successful small targets "membind" and T0065 are small helical proteins, and the larger successful targets 42, 56, and 61 are also essentially all helical, indicating that, at present, success is mainly limited to this type of structure.

CONCLUSIONS

In this article, we have explored only two of a range of possible measures. The number of examples is small, and the noise large, making it hard to reach statistically significant conclusions. Nevertheless, it is possible to make some observations. Model quality for most fold-recognition targets, measured by fraction of residues correctly aligned, does seem to have improved significantly from CASP1 to CASP2 but not from CASP2 to CASP3. The pattern is similar for both the very best predictions on each target and for the set of six best predictions. For the easier comparative models, alignment quality does not show any difference between the CASPs. Errors in this regime are dominated by the regions of the target that do not match the best available template closely, that is "loops," or secondary elements that have large relative shifts. Comparative modelers believe that improvements have been made in alignment quality, but in those portions of the structure that do superimpose well, not what is measured here. Quantitative evaluation of that type of alignment awaits a detailed study of progress in comparative modeling. Encouragingly, ab initio prediction on small targets does seem to have improved from CASP 1 and 2 to CASP3. Most success has been with all α structures, with good results on one of two targets in CASP2 (T0042 but not T0037), and two of two in CASP3. The best results are for a CASP3 mixed α/β target (T0077), suggesting progress with other architectures.

Fig. 4. Comparison of ab initio performance in the three CASPs. **A:** The best ab initio prediction on each target. (**B**) Average over up to six best predictions. The number of residues closer than 1, 2, 4, 8, and > 8Å to the equivalent residues in the target structure are shown in the bars. Targets are ordered by size, for each CASP. Yellow, CASP1; purple, CASP2; green, CASP3. Results are shown for the small targets, for which numerically intensive methods could be used. Although CASPs 1 and 2 have only one reasonably successful target (T0042), there are three in CASP3 (56, 61, and 77).

ACKNOWLEDGMENTS

This work was performed in part under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract W-7405-Eng-48.

REFERENCES

1. Panchenko A, Marchler-Bauer A, Bryant S. Threading with explicit models for evolutionary conservation of structure and sequence. Proteins Suppl 1999;3:133–140.
2. Sippl MJ, Lackner P, Koppensteiner WA, Domingues S. An attempt to analyze progress in fold recognition from CASP1 to CASP3. Proteins Suppl 1999;3:226–230.
3. Zemla A, Venclovas, C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins Suppl 1999;3:22–29.
4. Zemla A, Venclovas Č, Moult J, Fidelis, K. 1999. http://predictioncenter.llnl.gov/
5. Marchler-Bauer A, Bryant S. Measures of threading specificity and accuracy. Proteins Suppl 1997;1:74–82.
6. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
7. Zu-Kang F, Sippl MJ. Optimum superimposition of protein structures: ambiguities and implications. Folding Design 1996;1:123–132.
8. Zemla A, Venclovas Č, Moult J, Fidelis K. 1999. http://predictioncenter.llnl.gov/casp3/results/Dali-ProSup/
9. Holm L, Sander C. 1999. http://www2.ebi.ac.uk/dali/
10. Pearson WR. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics 1991;11:635–650.
11. Lackner P, Koppensteiner AW, Domingues F, Sippl MJ. Automated large scale evaluation of protein structure predictions. Proteins Suppl 1999;3:7–14.