Structure, Vol. 10, 435–440, March, 2002, ©2002 Elsevier Science Ltd. All rights reserved. PII S0969-2126(02)00731-1

Reliability of Assessment of Protein Structure Prediction Methods

Marc A. Marti-Renom,[1,4] M.S. Madhusudhan,[1,4] András Fiser,[1] Burkhard Rost,[2] and Andrej Sali[1,3]
[1]Laboratories of Molecular Biophysics
Pels Family Center for Biochemistry and Structural Biology
The Rockefeller University
New York, New York 10021
[2]CUBIC
Department of Biochemistry and Molecular Biophysics
Columbia University
New York, New York 10032

The reliability of ranking of protein structure modeling methods is assessed. The assessment is based on the parametric Student's t test and the nonparametric Wilcox signed rank test of statistical significance of the difference between paired samples. The approach is applied to the ranking of the comparative modeling methods tested at the fourth meeting on Critical Assessment of Techniques for Protein Structure Prediction (CASP). It is shown that the 14 CASP4 test sequences may not be sufficient to reliably distinguish between the top eight methods, given the model quality differences and their standard deviations. We suggest that CASP needs to be supplemented by an assessment of protein structure prediction methods that is automated, continuous in time, based on several criteria applied to a large number of models, and with quantitative statistical reliability assigned to each characterization.

Introduction

Protein structure prediction methods need to be ranked reliably by their quality. A reliable ranking helps developers to improve their approaches, and also enables users to apply the existing tools judiciously. Ranking of different methods generally consists of the following steps: (1) define a set of test sequences; (2) define one or more model quality criteria; (3) apply the modeling methods to the test sequences; (4) assess the methods by calculating the quality scores for the models; and (5) rank the methods based on a comparison of the corresponding model quality scores.

Here, we focus on an aspect of the fifth step above, the statistical reliability of ranking of two modeling methods. The reliability of ranking increases with the number of test sequences and the difference in quality between two compared methods. When the number of test sequences is not large or the quality difference between the two methods is relatively small, a question arises as to the statistical significance of the observed ranking. Does the ranking of the methods reflect their true performance or is it a consequence of chance? If the ranking is not reliable, how many more test sequences are needed for it to become reliable? Alternatively, how large should the difference in quality between the two methods be for the ranking to be reliable, given the available number of test sequences?

As an illustration, we assess the reliability of the ranking of comparative modeling methods at the fourth meeting on Critical Assessment of Techniques for Protein Structure Prediction (CASP) [1]. In the CASP experiments, models for protein sequences are calculated shortly before their actual structures become available. The models are first assessed objectively by a large number of numerical criteria [2]. These numerical criteria are then interpreted by an assessor who provides the final assessment and ranking of the participants' methods at a CASP meeting. It is generally assumed that the ranking at CASP is reasonably accurate [3–6] although concerns about its reliability due to a small number of test sequences have also been expressed [7, 8]; for example, only 14 test sequences were available for the comparative modeling category at CASP4. A quantitative analysis of the statistical significance of the ranking of the methods based on the CASP models has not been published yet. In this communication, we focus narrowly on the reliability of ranking of comparative modeling methods at CASP4 based on a single model quality criterion. Our results raise doubts about the ability to distinguish among the quality of predictions from the top comparative methods under the conditions of CASP4, contrary to some reviews [6], and suggest what is needed for reliable ranking in the future. This examination of the statistical significance of ranking of comparative methods at CASP4 is not a criticism of CASP as a whole; the CASP experiment has generally been recognized as a positive contribution to the field of protein structure prediction. We conclude by discussing some aspects of the assessment of protein structure prediction in general.

Methods

We define the statistical significance of the ranking between two modeling methods based on a single model quality criterion. While a comprehensive characterization of modeling methods usually requires multiple quality criteria, such as the fraction of the test sequence modeled, the accuracy of side chains, or the accuracy of loops, the test of statistical significance may be repeated independently for each criterion. Separate consideration of the individual model criteria is useful because different aspects of a structure may be predicted best by different methods.

We first propose one reasonable quality criterion for comparative models. Although we do not focus on model quality criteria [2, 8, 9], we need to define one such criterion

[3]Correspondence: sali@rockefeller.edu
[4]These authors contributed equally to this work.

Key words: comparative protein structure modeling; homology modeling; model assessment; statistical significance; CASP; EVA; LiveBench

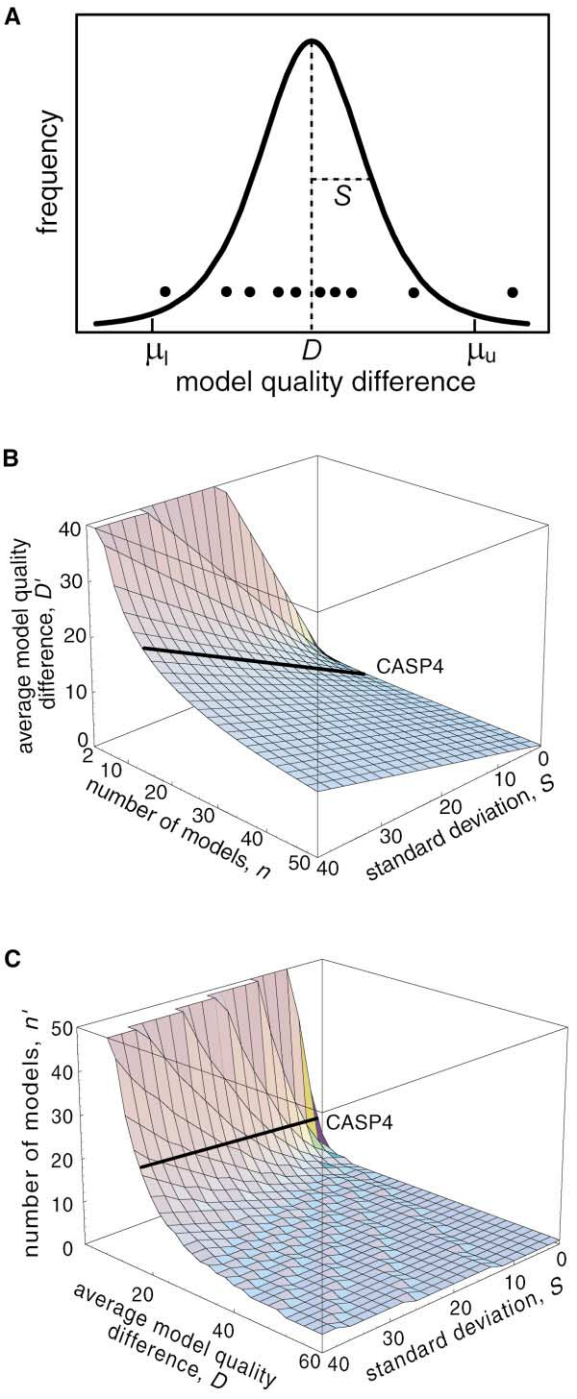

Figure 1. Explanation of Statistical Significance of Ranking Two Modeling Methods

(A) The black dots indicate the model quality differences for a hypothetical sample of common models produced by the two compared methods. The normal distribution of the model quality differences corresponding to the average and standard deviation of the sample is shown as a bell curve. The lower and upper bounds μ_l and μ_u on the average model quality difference for the whole population of all possible models, estimated from the sample as defined in the text, are indicated on the horizontal axis. *D* and *S* are sample average and standard deviation of model quality difference, respectively. When 0 lies between μ_l and μ_u, the difference between the performances of the two methods is not statistically significant at the given confidence level.

to be able to illustrate the issue of statistical significance of ranking with a practical example. Two fundamentally distinct features that are not comparable to each other describe the quality of a model: (1) the fraction of the protein sequence that is modeled (i.e., coverage) and (2) the accuracy of the modeled region. There are many criteria to assess the accuracy of different aspects of a model (e.g., core, loops, and side chains) [2, 8, 9]. In addition, the coverage and accuracy are generally dependent on each other. The smaller the modeled fraction of the sequence, the more accurate is the model; for example, the accuracy of a model can be increased at the expense of coverage by retaining only the core of the fold and eliminating loops and termini from the model. Hence, it is necessary to combine coverage and accuracy into a single number when comparing methods that model different parts of the test sequences, as is the case at CASP. While there is no single best way of defining accuracy or of combining coverage and accuracy, there are a number of reasonable combined criteria. We define model quality as the average percentage of the C_α atoms that are within 1, 2, and 3 Å of their correct positions upon least squares superposition of the model with the corresponding experimentally determined structure. These numbers were taken from the CASP4 evaluation site (http://predictioncenter.llnl.gov; January 31, 2001). The resulting quality criterion has a reasonable dynamic range in the sense that it can discriminate between backbones of models differing only in minor details (i.e., the 1 Å cutoff) as well as models differing at the level of alignment (i.e., the 3 Å cutoff). The selected quality criterion is closely related to perhaps the most frequently used single measure of model quality at recent CASP meetings: the "similarity" curves [10], the rmsd/coverage graphs [11], and the "global distance" curves (http://predictioncenter.llnl.gov); the criterion approximates the area under the global distance curve from 0 to 3 Å.

Prediction methods can be compared most reliably when they are tested under identical circumstances; for example, two modeling methods cannot be ranked by comparing the quality of an easy model based on a close template structure from one method with the quality of a difficult model based on a distant template structure from the other method. Thus, the best way to rank two methods is to assess their models for the same test sequences. Such a comparison is quantified by the distribution of the pairwise model quality differences, one difference for each of the common models (Figure 1).

In principle, there are two extreme possibilities: (1) the model quality difference is distributed around zero, indicating the lack of a statistically significant difference

(B) The average model quality difference *D'* that is required for statistical significance (95%) is plotted against the number of common models *n* and standard deviation of the observed model quality differences *S*. A line corresponding to 14 models is drawn to indicate the maximum possible number of common comparative models at CASP4. The actual average number of common models is only 8.5.
(C) The number of common models *n'* that is required for statistical significance is plotted against the average *D* and standard deviation *S* of the observed model quality difference.

between the performances of the two methods and (2) the distribution is shifted from zero significantly relative to its standard deviation, indicating a statistically significant difference between the performances of the two methods. If the distribution of the model quality difference is at least approximately Gaussian, the lower and upper bounds on the average model quality difference of the whole population of the models sampled by the common models in the test set are [12]:

$$\mu_{u,l} = D \pm \frac{t(n-1,c)S}{\sqrt{n}}$$

where D is the average model quality difference for the pairs of the common models in the sample, S is the standard deviation of the model quality difference in the sample, $t(n-1,c)$ is the Student's t distribution for an unknown population average and standard deviation [12], $n-1$ is the number of degrees of freedom in the sample (one less than the number of common models), and c is the confidence level of the bounds (95% in this paper). According to the paired samples Student's t test, a modeling method is significantly better than another at the given confidence level if the estimated interval of the average model quality difference lies below or above zero. Conversely, when the signs of the lower and upper bounds μ_l and μ_u differ, the performance of the two methods is not distinguishable at the confidence level of c.

It is important to distinguish between the magnitude of the average model quality difference and its statistical significance. A large average model quality difference can be statistically insignificant if the number of common models is small and the standard deviation of the model quality difference is large. A small number of common models and a large standard deviation of model quality difference require a large average model quality difference for confident ranking of two methods (Figure 1B). Similarly, a small average model quality difference and a large standard deviation of model quality difference require a large number of common models for confident ranking of two methods (Figure 1C).

The definitions above are sufficient to achieve the main aim of the study, to determine for each pair of methods whether it is possible to discriminate their performances based on a limited number of test sequences and the chosen model quality criterion. In addition, the pairwise method comparisons were used in two steps to construct an approximate ranking that lists all of the assessed methods. First, for each method i, a temporary ranking was constructed for it and for all other methods j with which it shared at least one model, using the average model quality differences. Second, the final ranking list was obtained by averaging the rank positions in all the temporary ranking lists. In general, any final ranking list will be frustrated in the sense that method i assessed to be better than method j in the direct pairwise comparison may be positioned worse than method j in the final ranking because of the impact of the pairwise comparisons with other methods, based on different sets of common models. While frustration can be quantified by the standard deviation of the temporary rankings, it is a good reason for focusing on the pairwise comparison of the methods.

Results

At CASP4, 123 methods were used to calculate models for at least one of the 14 comparative modeling test sequences (1,181 models in total; http://predictioncenter.llnl.gov/casp4; January 31, 2001). Only the 107 methods that were used to calculate models for at least two test sequences were analyzed further (1,165 models). When a method was used to calculate multiple models of the same test sequence, only the first model was considered. In total, there were 1,131 "first" models calculated by the 107 methods, corresponding to an average of 10.6 models per method. The average number of models common to all pairs of the 107 methods was 8.5.

Figure 2 shows the number of models calculated by each method, the statistical significance of ranking of each pair of methods at a confidence level of 95%, and the number of additional common models, if any, required for reliable pairwise ranking based on the model quality criterion defined in Methods. Upon visual inspection, the performance of the 107 methods falls into four weakly defined clusters.

The first cluster consists of the top eight methods (Figure 2). Their performance is not significantly different from each other according to the Student's t distribution statistics. Reliable ranking of the top eight methods is precluded by small numbers of common models (12.7 on average), small average model quality differences (1.3% on average), and large standard deviations of model quality difference (4.6% on average). Reliable ranking of the top eight methods would require on average 47 additional common models (Figure 2). In contrast, the 14 CASP4 test sequences are sufficient to rank the fictitious method with perfect alignments (method 000) better than any of the top eight methods; the average values of the average and standard deviation of model quality difference of method 000 with the top eight methods are 12.3% and 8.1%, respectively.

The first cluster of the top eight methods appears to be marginally distinguishable from the following $\sim$62 methods, which in turn are marginally distinguishable from the next $\sim$27 methods, followed by the least $\sim$10 accurate methods.

How reliable is the assessment of the statistical significance by the paired samples Student's t test? This parametric test is valid when the model quality difference is distributed approximately normally (it is not necessary that the model quality itself be distributed normally). When the model quality difference is not distributed normally, the estimated significance can be lower or higher than the actual significance. According to the Anderson-Darling test [13], the distribution of the model quality difference is normal for essentially all pairs of modeling methods (data not shown).

There are several other statistical methods suitable for assessing the significance of the average model quality difference. One such method is the paired samples Wilcox signed rank test, a nonparametric test of statistical significance that assumes a symmetric continuous distribution of the model quality difference [12]. We repeated all calculations with the Wilcox test (data not shown) and reached the same conclusions as with the Student's t test.

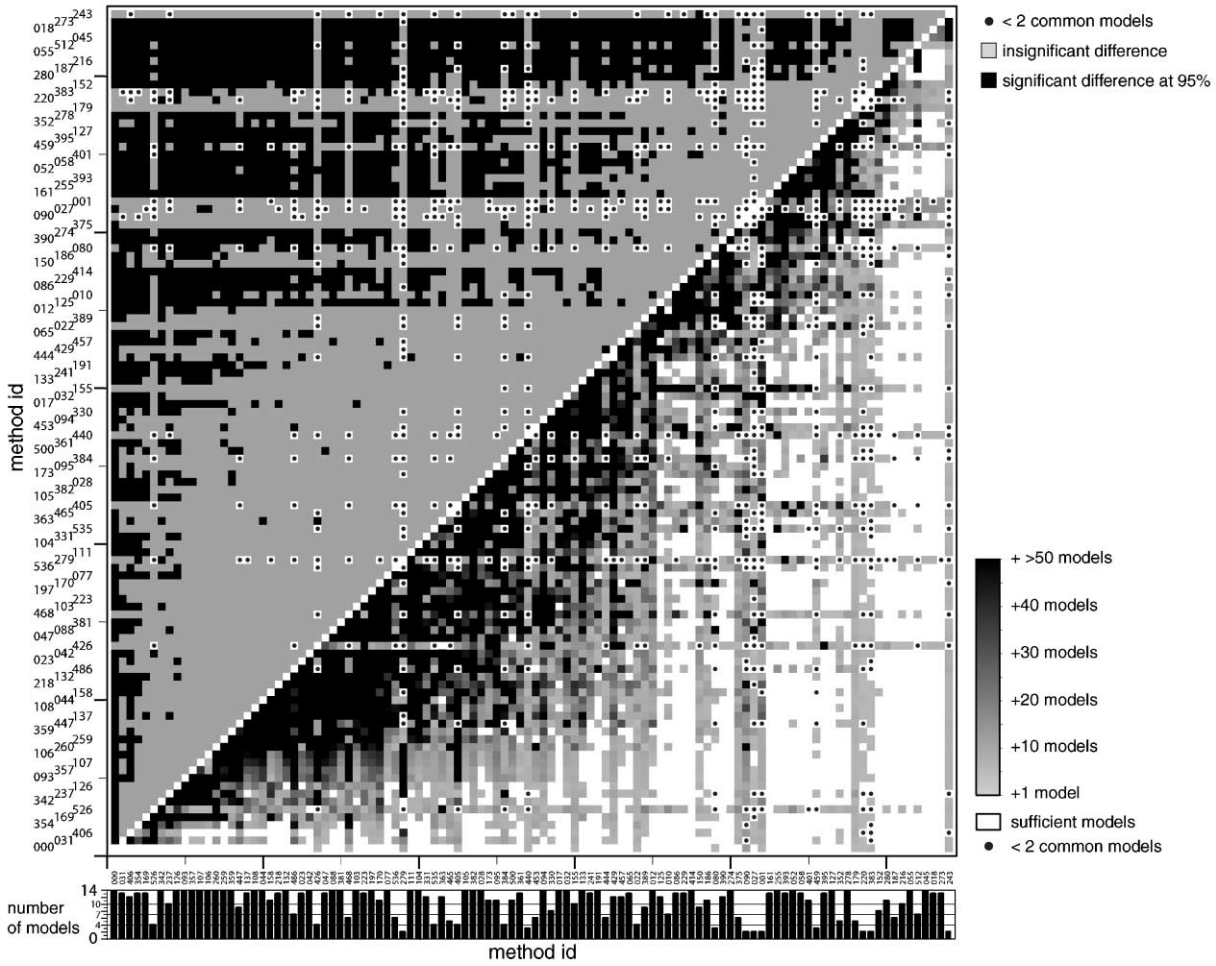

Figure 2. Comparison of Performances of Comparative Modeling Methods at CASP4, Based on the Model Quality Criterion as Defined in Methods

Upper diagonal (top legend): black and gray squares indicate pairs of methods whose performance is and is not statistically significantly different at the confidence level of 95%, respectively. Pairs of methods with less than two models in common, which could not be compared to each other, are indicated by black dots in white squares.

Lower diagonal (bottom legend): the intensity of gray indicates the number of additional common models that are needed to rank the two compared methods with statistical significance at a confidence level of 95%; if more than 100 additional models are needed, 101 is logged. The white squares correspond to pairs of methods that can already be ranked reliably based on a comparison of the common models submitted to CASP4. The histogram at the bottom shows the number of models calculated by each method. The method identifiers, as defined at CASP4, are given on the horizontal and vertical axes. Method 000 is a fictitious prediction method corresponding to models without alignment errors; a model consists of the C_α atoms of the closest template structure that is optimally aligned with the actual target structure. Method 000 is not used in the calculation of the ranking list.

The order of the methods in the overall ranking list is somewhat arbitrary. The uncertainty of a position in the overall ranking list can be measured by the standard deviation of the method's rank over all the temporary ranking lists (Methods). This frustration is an unavoidable consequence of different common sets of models for the different pairwise method comparisons. For example, the first method in the final ranking list is positioned from 1 to 22 in the temporary ranking lists, with the standard deviation of 4.7. The eighth method in the final ranking list is positioned from 2 to 31, with the standard deviation of 4.9. Similarly, the overlap between the clusters of methods is exemplified by the averages and standard deviations of temporary rankings of the ninth method (13.0 $\pm$ 7.3) and the 36th method (37.1 $\pm$ 18.1). We note again that, in contrast to the overall ranking list, the statistical significance of a pairwise method comparison is not arbitrary. Thus, ranking at CASP would be significantly simplified if all the methods were applied to all the test sequences.

Discussion

We describe a general procedure for quantifying reliability of ranking of two protein structure prediction methods, given a single model quality criterion. This procedure relies on the parametric Student's t test or the nonparametric Wilcox signed rank test of statistical significance of the difference between paired samples. Depending on the number of test models and the difference

between the performances of the two methods, it may or may not be possible to conclude that one method is more accurate than the other. Reliable ranking of different methods is important in the development as well as the use of the methods. This approach to ranking was illustrated by application to the ranking of comparative modeling methods at CASP4.

CASP meetings have become one of the most influential venues for assessing protein structure modeling methods [14, 15]. For example, it was reported without elaboration that team 126 retained its leading position in the comparative modeling category [6]. However, the performance of method 126 is statistically indistinguishable from that of 53 other methods with more than ten models, as judged by the model quality criterion and the statistical significance test described in this communication (Figure 2). Similar results apply to the other top seven methods. Thus, our results indicate that the number of test sequences at CASP4 may not be sufficient to distinguish reliably between the top eight comparative modeling methods, or, conversely, that the differences between the top eight methods may be too small for the methods to be distinguished based on the small number of test sequences available at CASP4.

The use of other single or multiple model quality criteria is not likely to rank any modeling method overall as significantly better than the other top methods. The reason is that the current model quality criterion is a good measure of the largest differences among the comparative models at CASP4, which result from errors in alignment and modeling of insertions. Similarly, other reasonable approaches to assessing statistical significance of ranking based on the CASP4 models are also not likely to detect reliable ranking among the top methods. This suggestion is substantiated by the agreement between the Student's t and Wilcox rank order tests, especially since their underlying assumptions seem to be met. Our attempt to rank comparative methods tested at CASP4 highlights the need for calculating the statistical significance of ranking to prevent misleading subjective judgment.

The Student's t distribution statistics allow us to compute what is required for reliable ranking of methods at future CASP meetings. Methods can be ranked reliably only if their performance is significantly different from each other. Such significance can be achieved by (1) a large number of common models (Figure 1C), (2) a large average model quality difference between two methods (Figure 1B), or (3) a small standard deviation of model quality difference. If the aim is to discriminate between two methods whose difference corresponds to the average model quality difference between the top eight methods at CASP4 ($1.3 \pm 4.6\%$), 52 test sequences are required (Figure 1C). Alternatively, to discriminate between two methods based on only 14 common models, the average model quality difference has to be at least twice as large as that at CASP4 (Figure 1B).

It is useful to assess prediction methods in terms of different kinds of modeling errors at different levels of difficulty. Such a stratified assessment is useful because different methods may perform best in different circumstances. For example, a certain method may be relatively successful at alignment when there is little sequence similarity, but unsuccessful at side chain modeling based on high-sequence similarity. The important errors in comparative modeling include errors in recognizing weak sequence structure similarities, alignment, modeling of insertions, rigid body shifts, distortions, and side chains, as well as mistakes in detecting errors in a model [16]. For a comprehensive characterization of modeling methods, hundreds of test sequences are needed.

Human predictors and assessors are not likely to be able to handle many more test sequences than at the past CASP meetings. Predictors only have a few months to generate their models, and an assessor only has about 2 months to examine approximately 1000 models calculated by 100 methods; a rigorous examination that goes beyond the use of a single model quality criterion must depend on consideration of tens of quantitative assessment criteria and visual inspection of each model. It appears that testing with hundreds of sequences can be achieved only by automating both the modeling and assessment methods. Although the CAFASP section of CASP [17] already evaluates automated prediction methods, this assessment is the same as that of the other models and is thus exposed to the same problems. While there is clearly a continued need for subjective but judicious examination of the successes and failures of protein structure prediction, the CASP experiments need to be supplemented by large-scale, automated, and continuous assessments, such as those implemented in the LiveBench (http://www.bioinfo.pl/LiveBench) [18] and EVA (http://cubic.bioc.columbia.edu/eva/) [8, 19] web servers for assessing automated protein structure prediction methods.

Acknowledgments

M.A.M.-R. and A.F. were Burroughs Wellcome Fund Postdoctoral Fellows, and are currently Rockefeller University Presidential and Charles Revson Foundation Postdoctoral Fellows, respectively. A.S. is an Irma T. Hirschl Trust Career Scientist. Support by the Merck Genome Research Institute (A.S.), Mathers Foundation (A.S.), and NIH (GM 54762 to A.S. and GM62413 to B.R.) is also acknowledged.

Received: June 25, 2001
Revised: October 16, 2001
Accepted: October 24, 2001

References

1. Moult, J., Pedersen, J.T., Judson, R., and Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. Proteins *23*, ii–v.
2. Venclovas, C., Zemla, A., Fidelis, K., and Moult, J. (1997). Criteria for evaluating protein structures derived from comparative modeling. Proteins Suppl. *1*, 7–13.
3. Dunbrack, R.L., Jr., Gerloff, D.L., Bower, M., Chen, X., Lichtarge, O., and Cohen, F.E. (1997). Meeting review: the second meeting on the Critical Assessment of Techniques for Protein Structure Prediction (CASP2), Asilomar, California, December 13–16, 1996. Fold. Des. *2*, R27–R42.
4. Koehl, P., and Levitt, M. (1999). A brighter future for protein structure prediction. Nat. Struct. Biol. *6*, 108–111.
5. Sternberg, M.J., Bates, P.A., Kelley, L.A., and MacCallum, R.M. (1999). Progress in protein structure prediction: assessment of CASP3. Curr. Opin. Struct. Biol. *9*, 368–373.
6. Murzin, A.G. (2001). Progress in protein structure prediction. Nat. Struct. Biol. *8*, 110–112.

7. Venclovas, C., Zemla, A., Fidelis, K., and Moult, J. (1999). Some measures of comparative performance in the three CASPs. Proteins Suppl. *3*, 231–237.
8. Eyrich, V., Marti-Renom, M.A., Przybylski, D., Fiser, A., Pazos, F., Valencia, A., Sali, A., and Rost, B. (2001). EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics *17*, 1242–1243.
9. Cristobal, S., Zemla, A., Fischer, D., Rychlewski, L., and Elofsson, A. (2001). A study of quality measures for protein threading models. BMC. Bioinformatics *2*, 5.
10. Sanchez, R., and Sali, A. (1997). Evaluation of comparative protein structure modeling by MODELLER-3. Proteins Suppl. *1*, 50–58.
11. Hubbard, T.J. (1999). RMS/coverage graphs: a qualitative method for comparing three- dimensional protein structure predictions. Proteins *37*, 15–21.
12. Rees, D. (1987). Foundation of Statistics (New York: Chapman and Hall).
13. Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. (1992). Numerical Recipes (Cambridge, MA: Cambridge University Press).
14. Sippl, M.J. (1999). Who solved the protein folding problem? Structure Fold. Des. *7*, R81–R83.
15. Fischer, D., Elofsson, A., and Rychlewski, L. (2000). The 2000 Olympic Games of protein structure prediction; fully automated programs are being evaluated vis-a-vis human teams in the protein structure prediction experiment CAFASP2. Protein Eng. *13*, 667–670.
16. Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. *29*, 291–325.
17. Bujnicki, J.M., Elofsson, A., Fischer, D., and Rychlewski, L. (2001). Structure prediction meta server. Bioinformatics *17*, 750–751.
18. Bujnicki, J.M., Elofsson, A., Fischer, D., and Rychlewski, L. (2001). LiveBench-1: continuous benchmarking of protein structure prediction servers. Protein Sci. *10*, 352–361.
19. Rost, B., and Eyrich, V. (2001). EVA: large-scale analysis of secondary structure prediction. Proteins *45*, 192–199.

Note Added in Proof

A reader may want to see a detailed description of assessment of comparative modeling methods at CASP4 by Tramontano et al. on pages 22–38 in the special issue 5 of Proteins that became available online on January 28, 2002.