PROTEINS: Structure, Function, and Genetics Suppl 5:13–21 (2001)
DOI 10.1002/prot.10052

Processing and Evaluation of Predictions in CASP4

Adam Zemla,[1] Česlovas Venclovas,[1§] John Moult,[2] and Krzysztof Fidelis[1*]
[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 **The Livermore Prediction Center conducted the target collection and prediction submission processes for Critical Assessment of Protein Structure Prediction (CASP4) and Critical Assessment of Fully Automated Structure Prediction Methods (CAFASP2). We have also evaluated all the submitted predictions using criteria and methods developed during the course of three previous CASP experiments and preparation for CASP4. We present an overview of the implemented system. Particular attention is paid to newly developed evaluation techniques and data presentation schemes. With the rapid increase in CASP participation and in the number of submitted predictions, special emphasis is placed on methods allowing reliable pre-classification of submissions and on techniques useful in automated evaluation of predictions. We also present an overview of our website, including target structures, predictions, and their evaluations (http://predictioncenter.llnl.gov). Proteins 2001;Suppl 5:13–21.** © 2002 Wiley-Liss, Inc.

Key words: protein structure prediction; evaluation methods; CASP4

INTRODUCTION

An outline of the Critical Assessment of Protein Structure Prediction (CASP) infrastructure implemented at the Livermore Prediction Center is presented. The main purpose is to provide an overview of the steps involved in automated assessment of predictions, beginning with format verification and ending with numerical and graphic presentation of the results. Many elements of this process carry over from previous CASPs, and thus we concentrate on the new developments, presenting a summary of the rest. The main tasks addressed at the Livermore Prediction Center are as follows:

1. *Prediction targets:* Target solicitation and collection from crystallographers and nuclear magnetic resonance (NMR) spectroscopists, including verification of sequence data and oversight of the target coordinate release status.
2. *Submission of predictions:* Format verification, submission updates, and verification of compliance with specific target deadlines.
3. *Evaluation of predictions:* (a) carrying out the necessary calculations; and (b) development of evaluation methods.
4. *Presentation of results:* Organization of the evaluation data, including generation of graphical summaries with links to detailed results and generation of an adjustable interface allowing for user-defined comparison of results.

In the area of results analysis, probably the most significant development since CASP3 was an increase in the need for overview-type presentations. With an almost threefold increase in the total number of submitted predictions, from approximately 3,800 in CASP3 to >11,000 in CASP4, it became practically impossible to analyze the results without the aid of summaries, which provide an uppermost organizational layer and a guide to any further comparisons. In this spirit, we have developed a number of graphic comparison tools designed to capture at least the overall quality of any given prediction. Beginning with these overviews, more detailed presentations of the results follow as a second layer and are accessible from the first through multiple HTML links.

The second significant development concerns the structure superposition method working in the sequence-independent mode. Until CASP3, we have used outside techniques,[1,2] recognizing the value of their track record in the community. However, with the large increase in the number of predictions submitted to CASP, it became prohibitively difficult to rely on methods that are not available locally. To remedy the situation, we have extended the GDT software[3] to include the local–global alignment package (LGA).[4] When tested on CASP3 data, on average, LGA produced slightly larger superimposed sets of atoms than the previously used methods when applying the same cutoff values. LGA was the major superposition engine for all CASP4 results generated at the Livermore Center.

SUBMISSION OF PREDICTIONS

All CASP4 and CAFASP2 predictions were received at the Livermore Prediction Center. Both high volume and the requirements of subsequent numerical evaluation

Grant sponsor: NIH; Grant number: LM07085-1; Grant sponsor: DOE; Grant number: DE-FG02-96ER 62271.

§Joint affiliation with Institute of Biotechnology, Graičiūno 8, 2028 Vilnius, Lithuania

*Correspondence to: Krzysztof Fidelis, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94551. E-mail: fidelis@llnl.gov

Received 19 July 2001; Accepted 29 October 2001

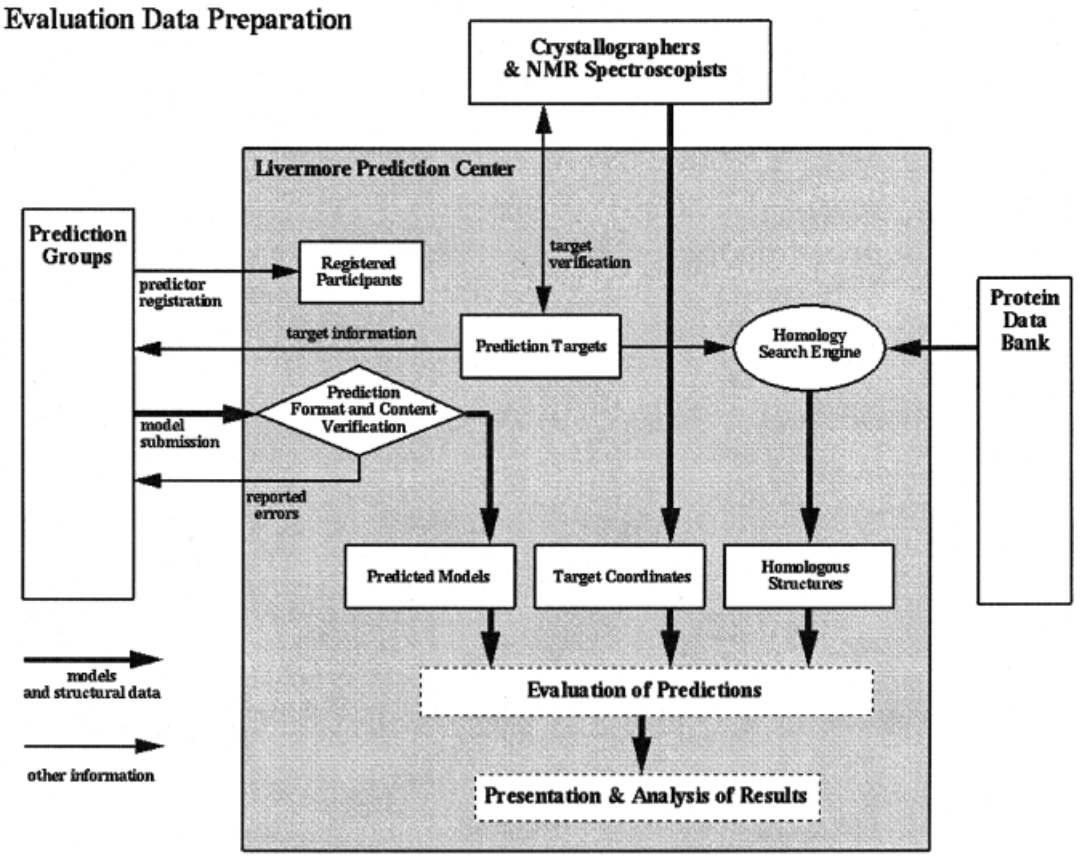

Fig. 1. Organization of the prediction submission process and preparation of target data.

placed high demands on format consistency and proper content of submissions. An automatic verification engine was used to ensure high standards of accepted data. The verification module was based on a standard UNIX sendmail utility, with the addition of Perl scripts to organize the flow of data, and programs written in C to handle the verification. Any submission format errors were quickly diagnosed and suggestions on how to amend them mailed back to predictors. Submissions were governed by the following set of rules:

1. All predictions were accepted electronically.
2. Each submission was automatically tested by the format verification server.
3. Models conforming to format and submission deadlines were assigned an accession code.
4. A unique accession code was composed of the following elements:
   i. Prediction target ID
   ii. Format category designator
   iii. Predictor group number
   iv. Model index (a number assigned by predictors to rank their submissions 1–5)
5. The following formats were used:
   i. TS (tertiary structure): predictions submitted in the form of atomic coordinates (three-dimensional [3D] models)
   ii. AL (alignment): predictions submitted in the form of sequence alignments to publicly available structures
   iii. SS (secondary structure): assignments of secondary structure to target protein sequences
   iv. RR (residue–residue contacts): predictions submitted in the form of Cβ–Cβ distances
6. Up to five models were accepted from each prediction group on any given target; primary attention was paid to only one of the models (designated by the predicting group as model index 1).
7. Submission of a duplicate model (same target, group, model index) replaced a previously accepted model, provided it was received before target's prediction deadline.

Figure 1 shows a schematic of the submission process, including target coordinates and homology data preparation.

ORGANIZATION OF THE EVALUATION SYSTEM

In CASP, evaluation is performed by comparison with experimental structures, which for each target protein define the "standard of truth." The Livermore Prediction Center coordinated the information for each of the prediction targets (i.e., sequence, source, advancement of struc-

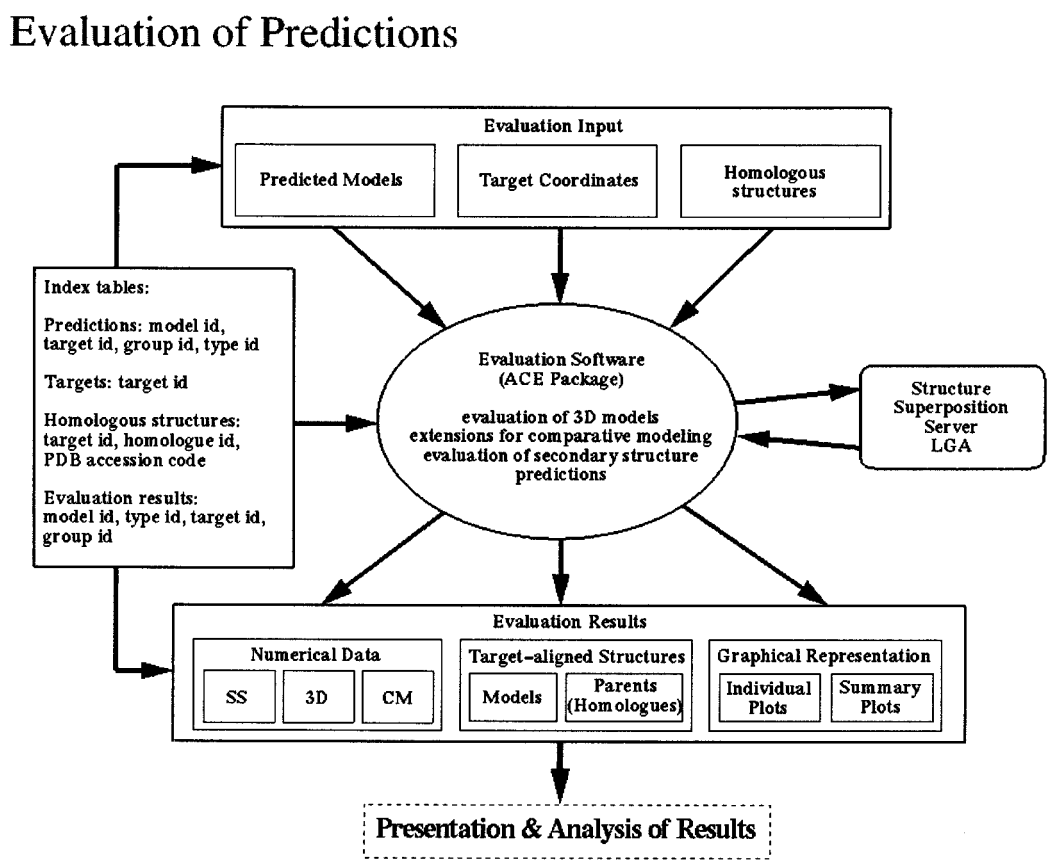

Fig. 2. Organization of the prediction evaluation system.

ture solution). These data were made available to predictors via a web interface.

As soon as the experimental structures were released, we collected information on their structural homologues available at the time. With approximately 50 new structures deposited to the Protein Data Bank (PDB) every week, it was important to capture the release status of these homology-related data at closing of the prediction time window for each of the targets. Homologous structures were also needed in the subsequent evaluation of submitted models. In comparative modeling, one or more of the closely related parent structures (modeling templates) were identified for this purpose. In fold recognition, more extensive lists of target-related structures were compiled, together with the corresponding levels of structural similarity they shared with target proteins. In both cases, we have used the ProSup structure similarity search procedure as provided by Sippl's group.[2,5] In comparative modeling, final selection of the principal parent structure involved further careful examination of the similarity between parent and target structures, using the LGA method.[4]

Predictions were evaluated as (1) general 3D models submitted in either the TS or AL formats and typically generated by methods in the ab initio (new fold) or fold recognition categories; (2) high-quality 3D models (denoted CM) typically generated by comparative modeling; and (3) assignments of secondary structure (denoted SS). Further division of the prediction targets into evaluation categories was addressed by the CASP4 independent assessors and is more broadly discussed in two articles included in this issue.[6,7] Residue–residue contact predictions were not evaluated at the Livermore Prediction Center. Hubbard's evaluation of these results is available at http://predict.sanger.ac.uk/casp4/. After the end of the CASP4 prediction season, all the submitted models, coordinates of the corresponding target structures, and data on relevant related proteins were assembled as the evaluation input files. Index tables containing model, target, prediction type, and parent structure identification labels, as well as the PDB accession codes, were used for quick reference by the evaluation program package (ACE), and by the data presentation engines (Fig. 2). Access to more than 14,000 3D models and 2,000 other predictions and their evaluation data is provided through our website.

EVALUATION OF 3D MODELS

In our presentation of CASP results, we have used both sequence-dependent and sequence-independent methods of comparing model with target. The sequence-dependent mode optimizes superposition of the two structures under the assumption of strict 1:1 correspondence between their residues. The sequence-independent mode forgoes that restriction. In cases in which the fold is correctly identified, but the target sequence alignment with the template structure is poor, the sequence-independent method will

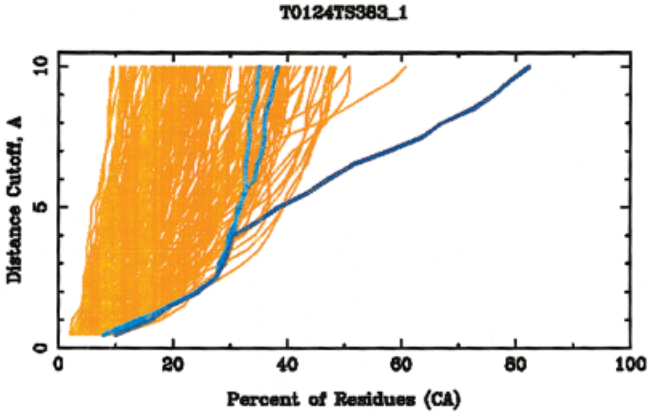

Fig. 3. Overall prediction quality graph for all models submitted on CASP4 target T0124 derived from multiple rigid body sequence-dependent superpositions of model and target structures. Each line corresponds to one predicted structure and shows percentages of the model fitting under the corresponding Cα–Cα distance cutoffs. Models submitted by one particular prediction group (383) are shown in blue (model 1) and cyan (other models).

allow detection of the incorrectly aligned regions in a two-step process: (1) a best possible superposition between model and target is obtained, and (2) errors in the relative sequence alignment are calculated on the basis of that superposition. This section describes the types of results calculated for the 3D model predictions.

Sequence-Dependent Method

Overall Prediction Quality Graphs (GDT)

The global distance test (GDT) summary graphs provide an approximate sorting of predictions by quality and a good starting point for further analysis (Fig. 3). Conceptually they are a variation of the root-mean-square deviation (RMSD)/coverage plots first introduced by Hubbard[8] but use distance rather than RMSD cutoffs. These plots consist of points identifying subsets of structure that can be fitted under a specified distance cutoff. In general, the more horizontal the curve corresponding to a particular model, the better the prediction. In the HTML presentation, clicking on a line identifies specific prediction along with other predictions submitted by the same group (blue and cyan, respectively). At this point, additional links provide a comparison of secondary structure assignments in the target and model, and the longest continuous segment (LCS) analysis. These plots identify precisely the LCS in the model structure that do not deviate from the target by more than a specified Cα RMSD. As in the LCS plots, results of the GDT analysis may also be displayed for specific models, with local prediction quality plotted as a function of position in the sequence. With this type of GDT plot, similarity between predicted and experimental structures may be assessed over regions that are not necessarily continuous in sequence. Both specialized techniques were described previously,[3] and thus we do not provide specific graphical examples here. Viewing of the 3D models and of the corresponding experimental structures is made available via the RASMOL graphic package, written by Sayle.[9]

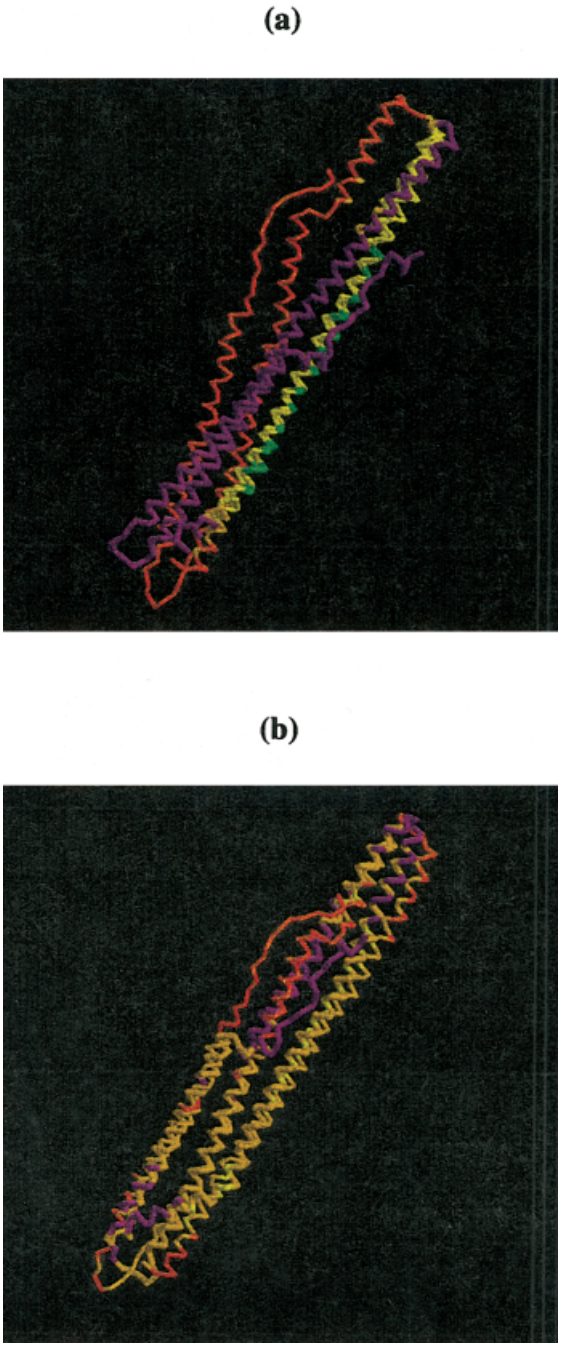

Fig. 4. Comparison of sequence-dependent superpositions obtained between predicted (group 383, model 1) and target structures of T0124. **a:** Superposition calculated with a lower cutoff (4.0 Å) identifies correct prediction of only one helix. **b:** For the same prediction, superposition calculated with a higher cutoff (8.0 Å) identifies the overall structure similarity. In both a and b, residues deviating by $<$2, 4, and 8 Å are shown in green, yellow, and orange, respectively. Segments of model and target not superimposed are shown in red and violet, respectively.

Links to numerical data on the quality of predictions are also provided. In particular, the GDT_TS (total scores) measure provides a reasonable single-value approximation of the tertiary structure prediction quality. The GDT_TS is defined as an average of four separate GDT

calculations identifying maximal sets of residues at 1, 2, 4, and 8 Å distance cutoffs.

The GDT summary assessment shown in Figure 3 highlights the risk of using single cutoff values in generating overall summaries. Specifically, the kink in the blue line represents a transition between a family of structural superpositions identified as optimal for distance cutoff values of approximately <4 Å, and another family compatible with cutoffs of >4 Å. In the first case, only a single helix is identified as structurally aligned with the target, while in the second case, similarity extends essentially over the entire structure [Fig. 4 (a) and (b), respectively].

Cα–Cα Deviation Stripcharts

Another means of quickly comparing all predictions on a given target are the Cα–Cα deviation strip charts. This specific representation is generated based on the best model-target rigid body superposition identified during the course of the GDT analysis using a 4-Å distance cutoff. With the caveat of the risk involved with using superpositions obtained under single cutoff values, discussed in the previous section, this approach helps identify specific regions in a prediction that are correctly modeled [Fig. 5(a)]. In the HTML presentation, each stripe provides a link to a RASMOL rendering of the 3D superposition of model and target structures [Fig. 5(b)].

Sequence-Independent Method

Sequence-dependent superposition methods are unable to identify regions of structural similarity in a prediction that are not correctly aligned by sequence. Sequence-independent methods will identify such regions and provide a direct measure of alignment accuracy. The LGA algorithm[4] is now used for all such evaluations at the Livermore Prediction Center.

We have implemented a calculation of the overall alignment quality [Fig. 6(a)], which also permits sorting by either exact or more relaxed criteria of alignment correctness (sorting allowing ±4 residue shift is shown). The corresponding strip chart [Fig. 6(b)] shows regions of a model that are correctly aligned. Serving as top-level overviews, these two graphs also provide HTML links to 3D representations of superimposed target and model structures [Fig. 6(c) and (d)].

The example shown in Figure 6(d) demonstrates how an essentially correct prediction of structure, although completely misaligned in sequence, is still identified by the sequence-independent superposition. Identification of this similarity is not possible in the sequence-dependent regime.

Outline of Previously Developed Evaluation Methods

Evaluation criteria have been designed to address different aspects of prediction, features that are particularly difficult to model, and characteristics relevant to protein function. They were also designed to single out elements of protein structure and to eliminate the effect of possible experimental uncertainties. Additional criteria assess specific stages in the process of prediction, success of the

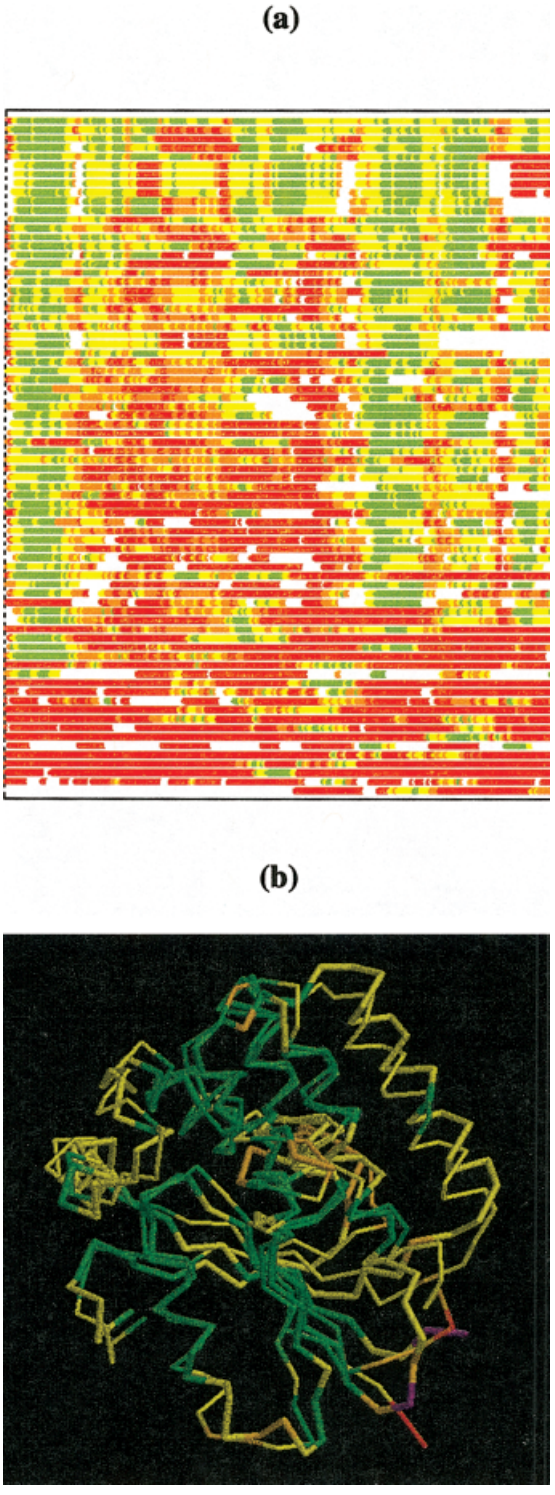

Fig. 5. **a:** Cα–Cα deviation strip chart for all predictions on CASP4 target T0117 derived from a single rigid-body sequence-dependent superposition of model and target structures. Each stripe corresponds to a single prediction shown as a function of sequence. Residues superimposed within 2, 4, and 8 Å are shown in green, yellow, and orange, respectively. Residues with Cα–Cα deviation of >8 Å are shown in red, and those not predicted in white. **b:** RASMOL rendering of the model (thin) and target structures (group 31, model 1, first stripe of the chart in a). The color scheme is the same as in the strip chart for both model and target, except for target residues that do not superimpose with model (violet).

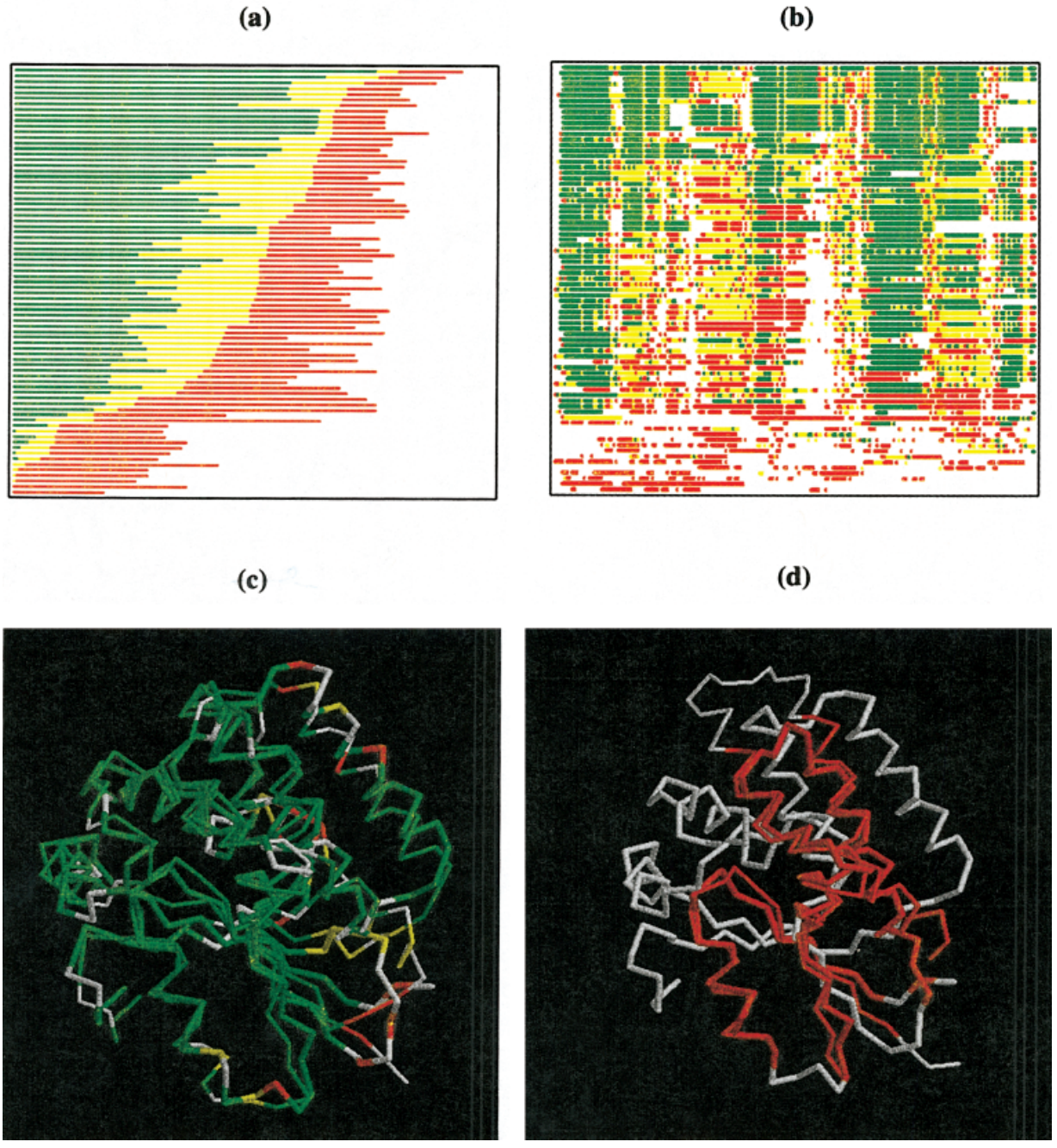

Fig. 6. **a:** Overall alignment quality bar graph for all predictions on CASP4 target T0117 derived from a single rigid-body sequence-independent superposition of model and target structures. Residues with correct sequence alignment (%) are shown in green and those aligned within ±4 residues in yellow. Residues superimposed structurally, but not aligned by sequence, are shown in red and the remainder, including those not predicted in white. **b:** Alignment quality strip chart for predictions shown in a: alignment quality plotted as function of position in sequence. Color scheme as in a. **c:** RASMOL rendering of the best model (thin) submitted on this target (group 31, model 1, first stripe of the chart in b). Colors correspond to the strip chart representation. **d:** RASMOL rendering of a model capturing fold similarity but failing to align residues correctly by sequence (group 186, model 1, stripe second from the bottom in b). Colors correspond to the strip chart representation.

refinement techniques, and accuracy of the model reliability estimates. A more extensive overview is provided in refs. 3, 10, and 11.

Basic Measures

The RMSD between model and target is used to measure differences between atomic coordinates, with results dependent on structural superposition, and between dihedral angles, independent of superposition. For coordinates, results are calculated for all atomic positions or subsets, including Cα, main-chain, and side-chain atoms. RMSDs over dihedral angles are calculated separately for ϕ/ψ and for χ angles. Completeness of a prediction determines how many atomic positions or dihedral angles could be included

in the evaluation, and these numbers are provided for each submission. The following subsets of structure are used in the general assessment of 3D models: (1) residues of the secondary structure elements; (2) amino acids that are on the surface of a protein and those that are buried; and (3) residues not affected by possible experimental uncertainty, such as disorder or crystal contacts.

Additional Measures for High-Quality Models

To evaluate comparative modeling predictions, additional criteria have been developed. In the design considerations, particular attention was paid to the parts of the target structure that differ from any of the homologues, to the correct selection of the parent structure, and relevance to protein function. The resulting additional subsets include (1) angles that have a different rotameric assignment in target and parent structures; (2) chain segments that have moved significantly relative to the parent structure; (3) segments of the target structure for which selection of a parent other than the one closest by sequence is preferred; (4) "core" and "loop" segments; and (5) regions of structure that are in direct contact with ligand molecules.

RMSD Details of Loops

Loops as difficult to predict regions of structure were defined based on global (i.e., LGA) superposition of target and parent structures. Corresponding residues with Cα distances greater than cutoff (2.5 Å) were assigned to loop segments. If fewer than three residues exist between such segments, they are merged together. To address modeling performance specifically on individual loops, Cartesian RMSDs in both global and local superposition are calculated on Cα, main-chain, and all atoms for each loop that contains at least three residues.

Evaluation of Model Refinement and Confidence Assessments

To evaluate the success of the refinement procedures, such as energy minimization or molecular dynamics, parallel submissions of unrefined and refined models were accepted and fully assessed. Estimates of position-specific reliability of submitted models were assessed as previously.[3]

EVALUATION OF SECONDARY STRUCTURE PREDICTIONS

For the evaluation of secondary structure predictions, we have used a similar overview approach as in the case of 3D models. Bar graphs showing success rates and percentage of predicted structure permit rapid comparison of all predictions submitted on a given target [Fig. 7(a)]. Each evaluated prediction links to a strip diagram comparing predicted and target structure secondary structure assignments. Strip charts comparing all predicted assignments on a given target are also available [Fig. 7(b)].

ORGANIZATION OF THE WEBSITE

The http://predictioncenter.llnl.gov website provides general information about the prediction experiment and

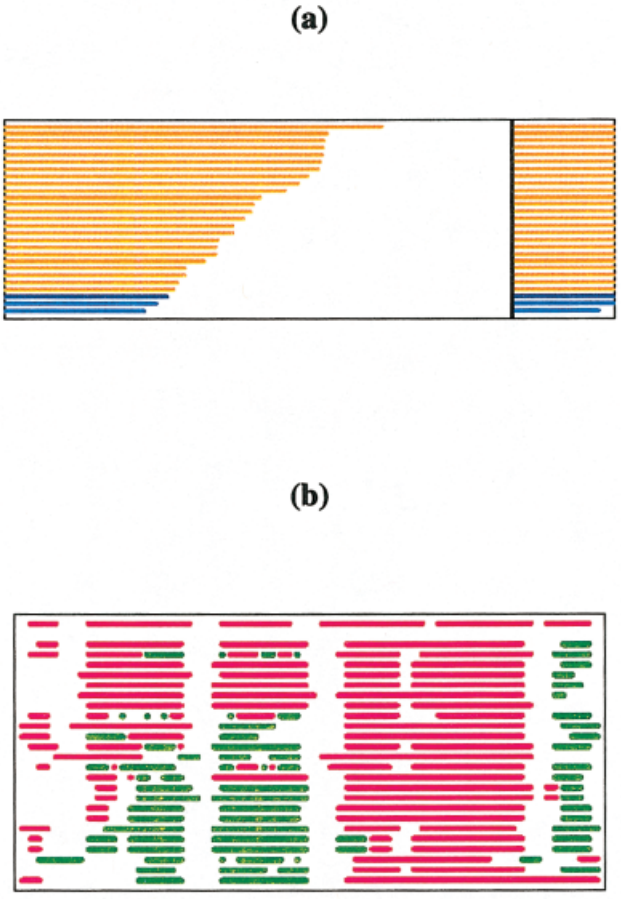

Fig. 7. Evaluation of secondary structure predictions for CASP4 target T0102. **a:** Bar graph showing prediction success in terms of the SOV measure[13] (%, left side of the graph) and fraction of predicted residues (%, at right). The same type of rendering is used for the Q3 results. **b:** Strip chart of predicted secondary structure assignments. First stripe from the top represents secondary structure assignment in the target structure, the remaining ones in all model 1 predictions. Color scheme: purple, helix; green, strand; white, coil; black, residues not predicted. This plot shows that only a few groups were able to predict the second and third helices correctly.

comprehensive access to prediction targets, original predictions, and evaluation results. The site also allows access to the visualization tools described above. Data for all four CASP prediction experiments are available. The website provides three main modes of access to evaluation data:

1. Summary graphics organized by prediction target, allowing quick comparisons of all predictions submitted on a given target structure, with four types of graphs available:
   a. Alignment accuracy plots (sequence-independent analysis of 3D models)
   b. GDT plots (sequence-dependent analysis of 3D models)
   c. Target-model Cα–Cα deviation summary plots (sequence-dependent analysis of 3D models)
   d. Sov and Q3 (secondary structure prediction evaluation results)

2. Interactive browsers, with user-defined comparison tables with links to graphic representations allowing for the selection of
   a. Prediction experiment (CASP1, 2, 3, or 4)
   b. Category of data
      i. 3D models (ab initio/new fold and fold recognition)
      ii. Extended evaluation for comparative modeling
      iii. Evaluation of secondary structure predictions
   c. Target proteins
   d. Prediction groups
   e. Evaluation criteria
   f. Subsets of structure

Generated results tables then provide links to graphic tools as follows:

   a. Summary plots of GDT analysis of all models for selected target
   b. Plots of largest superpositions between predicted and target structures (detailed GDT analysis of each prediction)
   c. Plots of LCS for each prediction
   d. Šali plots
   e. RASMOL plots of predicted and target structures in
      ii. Cα sequence-dependent superposition
      iii. GDT sequence-dependent superposition (only Cα atom pairs closer than 4.0 Å are used to obtain the superposition)
      iv. LGA sequence-independent superposition
   f. RASMOL plots of superimposed target and parent structures (for comparative modeling targets)
   g. Comparisons of residue-by-residue secondary structure assignments between predicted and target structures.

To simplify navigation through evaluation results from interactive browsers, a default set of measures and subsets of structure are also provided to generate the comparison tables.

3. Links to results generated by other evaluation methods:
   a. CASP4 fold recognition independent assessment (Sippl's assessment of fold recognition)
   b. Evaluation of models submitted for CAFASP2 experiment (results of the CAFASP assessment of fully automated predictions).

DISCUSSION

At the first CASP experiment in 1994, evaluation of a total of approximately 100 predictions presented a considerable assessment task. By CASP4, the number of submitted predictions increased more than 100-fold. The key elements allowing keeping up with this rapid increase were consistent, electronically readable formats, highly developed evaluation criteria, and considerable automation of the evaluation process. With the large number of predictions, classification by prediction quality became indispensable. In CASP, these techniques include numerical evaluations of the model–target structure similarity, implemented by Sippl's group[5]; RMS/coverage graphs, introduced by Hubbard[8]; and the methods developed by the Livermore Prediction Center. A subset are the representations of prediction quality as a function of position along the sequence. While still permitting visualization of all predictions on a given target with a single glance, they also permit comparison of prediction success in different regions of the model. Even if only approximate, these methods permit selection of models for further, more complete, assessment.

Several issues remain, including automated classification of protein architecture. An interesting example was provided in CASP4 by a model containing unlikely "knots" (prediction T0103TS218_1), a feature not easily detected by present numerical assessments. These rare topologic features have recently been studied more broadly,[12] including the development of an algorithm to detect them automatically. A discussion of this issue will undoubtedly be part of preparations for CASP5.

Nevertheless, most of the evaluation methods are now stable. There are two principal tasks ahead. The first task is full automation of evaluation, such that a prediction on any structure can be submitted at any time, and the results generated and returned without human intervention. This system will allow the methods to be used in the broader set of evaluation scenarios that are now emerging. The second task is the development of a system for classifying prediction methods. Such a system would provide cataloging of the existing techniques and remain open for the introduction of new ones. It would also provide a hierarchical organization to identify variations of the applied methodologies. For example, for a particular comparative modeling approach, it would specify what class of alignment algorithm was used. Over time, by providing a track record for all classified modeling methods, such a system will facilitate the development of techniques of assigning accuracy estimates to models obtained in real-life applications.

ACKNOWLEDGMENTS

This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract W-7405-Eng-48. This work was also supported by NIH grant LM07085-01 (to K.F.) and DOE grant DE-FG02-96ER 62271 (to J.M.).

REFERENCES

1. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
2. Feng ZK, Sippl MJ. Optimum superimposition of protein structures: ambiguities and implications. Fold Des 1996;1:123–132.
3. Zemla A, Venclovas Č, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins 1999;Suppl 3:22–29.
4. Zemla A. LGA program: a method for finding 3-D similarities in protein structures. 2000; accessed at http://predictioncenter.llnl.gov/local/lga/.
5. Lackner P, Koppensteiner WA, Domingues FS, Sippl MJ. Automated large scale evaluation of protein structure predictions. Proteins 1999;Suppl 3:7–14.

6. Tramontano A, Laplae R, Morea V. Analysis and assessment of comparative modeling prediction in CASP4. Proteins 2001; Suppl 5:22–38.
7. Sippl MJ, Domingues FS, Prlic A, Malik R, Andreeva A, Wiederstein M. Assessment of the CASP4 fold recognition category, Proteins 2001; Suppl 5:55–67.
8. Hubbard TJ. RMS/coverage graphs: a qualitative method for comparing three-dimensional protein structure predictions. Proteins 1999;Suppl 3:15–21.
9. Sayle RA, Milner-White EJ. RASMOL: biomolecular graphics for all. Trends Biochem Sci 1995;20:374–376.
10. Venclovas Č, Zemla A, Fidelis K, Moult J. Criteria for evaluating protein structures derived from comparative modeling. Proteins 1997;Suppl 1:7–13.
11. Zemla A, Venclovas Č, Reinhardt A, Fidelis K, Hubbard TJ. Numerical criteria for the evaluation of ab initio predictions of protein structure. Proteins 1997;Suppl 1:140–150.
12. Taylor WR. A deeply knotted protein structure and how it might fold. Nature 2000;406:916–919.
13. Zemla A, Venclovas Č, Fidelis K, Rost B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins 1999;34:220–223.