PROTEINS: Structure, Function, and Genetics Suppl 3:7–14 (1999)

NUMERICAL EVALUATION METHODS

Automated Large Scale Evaluation of Protein Structure Predictions

Peter Lackner, Walter A. Koppensteiner, Francisco S. Domingues, and Manfred J. Sippl*
Center for Applied Molecular Engineering, Institute for Chemistry and Biochemistry, University of Salzburg, Salzburg, Austria

ABSTRACT **Evaluation and assessment are critical issues in CASP experiments. Automated procedures are necessary to compare a large number of predictions with the target folds. The evaluation has to reveal the maximum extent of similarity between predictions and targets, it should be applicable across prediction categories, and it should be transparent and accessible to a wide community. Here we present an automated evaluation scheme which is an attempt to meet these requirements. In the implementation and execution of this scheme we had to solve or circumvent problems of convergence, where algorithms fail to find optimum solutions, problems of ambiguity where no unique optimum solution exists, and problems in ranking and interpretation. Key features of this implementation are (1) the root mean square deviation of structure superimposition is kept close to a constant value throughout the evaluation and (2) all structural matches found between two folds are taken into account. We discuss these points in detail and describe the numerical criteria used in the CASP3 evaluation. Proteins Suppl 1999;3:7–14.** © 1999 Wiley-Liss, Inc.

Key words: CASP; structure comparison; structure similarity; alternative alignments

INTRODUCTION

Proper evaluation and assessment are most sensitive issues in CASP experiments. In particular, the techniques used in the evaluations must reveal the maximum extent of similarity between prediction and target fold, they should provide a suitable basis for assessment, and they have to be transparent and accessible to a wide audience. The evaluation of CASP1[1–3] and CASP2[4–6] uncovered several difficulties associated with the large scale prediction analysis that go beyond the problems generally encountered in structure comparison applications.

These problems were addressed by some participants of previous CASP experiments at a FEBS prediction course organized by Anna Tramontano and Tim Hubbard at the IRBM in Pomezia.[7] The discussion resulted in many ideas and suggestions and the complexity of the problem was acknowledged. However, a general consensus on which methods might work could not be reached. On the basis of this discussion we decided that the evaluation criteria should:

- Consist of a minimum but sufficient number of numerical measures to highlight the important features of a prediction;
- Be applicable across all prediction categories;
- Take into account all structural matches that exist between predicted model and target fold;
- Resolve ambiguities in favor of the predictor; and
- Be transparent to a wide audience.

In addition to these goals several specific issues like the exhaustive determination of similarity between target folds and structures in PDB, the ranking of predictions, and the accuracy of sequence structure alignment in fold recognition, should be addressed.

The challenge was to implement an evaluation scheme that meets these requirements. The implementation reported here is based on rigid body superimposition as implemented in ProSup.[8,9] It was clear at the outset, that problems in structure comparison are varied and complex where a single approach might not be able to handle all subtleties. Therefore, we concentrated on potential problems and on the description of essential aspects of predictions. In the CASP3 evaluation we were confronted with expected and unexpected difficulties in protein structure comparison which we describe below.

PROBLEMS IN PROTEIN STRUCTURE COMPARISON

Structure comparison programs are often built on the principle that a suitable scoring function can be defined whose optimum corresponds to the most significant structural match. Other solutions are disregarded as subopti-

Grant sponsor: Fonds zur Förderung der Wissenschaflichen Forschung; Grant numbers: P11601-GEN and P11205-MOB; Grant sponsor: Fundação para a Ciência e a Tecnologia; Grant number: PRAXIS XX1/BD/4528/94.

*Correspondence to: Manfred J. Sippl, Center for Applied Molecular Engineering, Institute for Chemistry and Biochemistry, University of Salzburg, Jakob Haringer Straße 3, A-5020 Salzburg, Austria.
E-mail: sippl@came.sbg.ac.at

Received 12 March 1999; Accepted 7 June 1999

© 1999 WILEY-LISS, INC.

mal. The implicit assumptions are that the program always finds the optimum solution, and that this optimum is unique. We identified three classes of problems that originate from these assumptions:

- Failure to find the best solution due to numerical problems;
- Problems of proper ranking due to ambiguous or imprecise definition of structural similarity;
- Inability to detect structural matches due to the neglect of alternative solutions.

Problems of Convergence

Situations where a program fails to find the optimum solution are difficult to recognize. Obviously, such problems can have devastating consequences in large scale evaluations. We paid particular attention to this problem. A good strategy to detect problems of convergence is to reverse the order of comparisons. Superimposition of fold A and B should yield the same result as B and A. Distinct results usually imply failure of convergence. We used this principle in tuning the convergence properties of ProSup. A proper account of the technical details is beyond the scope of this presentation (see Lackner et al.[9]).

Ranking by Structural Similarity

A problem often encountered in structure comparison is to rank a set of folds F_i by their similarity to a target fold A. Examples relevant for CASP evaluations are the ranking of PDB folds with respect to the similarity to a target, or the ranking of predicted structures relative to a target. Ranking requires that structural similarity is expressed as an order relationship $S(A, F_i) \leq S(A, F_j)$, implying that $S(A, B)$ has to be a scalar quantity. Many established comparison techniques define structural similarity by two numbers, the root mean square (rms) error of superimposition together with the number of equivalent residues resulting in a vectorial quantity unsuitable for ranking. DALI solves this problem by combining several numbers to a single quantity, called z-score. The score is used for ranking and is also supposed to indicate the significance of a match.

Still the ranking obtained is often difficult to interpret. The problem originates from the construction of the underlying algorithms rather than the combination of numbers. Several established structure comparison techniques optimize the rms error of superimposition and the number of equivalent residues or related quantities. However, it is impossible to satisfy both simultaneously, since one quantity can be optimized on the expense of the other. A natural refinement is to maximize the number of equivalent residues while rms is kept close to a constant value. Control of rms values is achieved indirectly by a proper combination of ProSup parameters (see Lackner et al.[9])

To highlight this problem the example in Table I shows the ranking of folds in PDB relative to target T0079, the transcriptional activator MarA protein (1bl0[10]), obtained from DALI and ProSup, where the latter forces rms values to be smaller than 3.0 Å. For this target Vast reports only 1crx as a related structure. In the DALI table folds of comparable z-scores can vary widely in their respective rms values and number of equivalent residues (compare 2fok-A and 1aih-A, or 2ezk and 1vin, for example) and it is difficult to understand why these folds yield similar scores. In contrast the ProSup ranking is rather easy to interpret since the number of equivalent residues relates to an intuitive conception of structural similarity.

TABLE I. Database Search of the Transcriptional Activator MarA (1bl0) With ProSup and DALI†

ProSup			DALI			
PDB	L_1	R_1	PDB	L_1	R_1	z-score
1bl0.A	116	0.0	1bl0.A	116	0.0	24.9
1a36.A	82	2.7	1crx.A	88	3.3	4.0
1crx.A	81	2.5	1a36.A	99	3.1	3.9
1aih.A	71	2.3	2fok.A	65	6.5	3.4
1mty.B	68	2.9	1aih.A	84	3.6	3.4
1mty.D	67	2.8	1jhg.A	68	6.7	3.3
1a31.A	67	2.9	2ezk	64	7.1	3.0
1mhy.D	65	2.8	1vin	63	3.6	2.8
1ae9.A	64	2.4	1sly	69	4.1	2.8
1a3w.A	59	3.0	1ae9.A	66	2.3	2.7
1cfr	57	2.8	1aoy	60	4.4	2.6
1gcb	56	2.5	1pdn.C	55	3.4	2.6
1ova.A	56	2.7	1fip.A	59	5.5	2.6
2tct	56	2.8	1a04.A	51	3.5	2.5
1pkm	56	2.9	1bia	57	3.2	2.4
1phn.B	56	3.1	1kfs.A	57	5.0	2.4
1cpc.B	55	2.4	2tct	54	4.0	2.4
1pfr.A	55	2.4	2bby	58	3.5	2.4
1fok.A	55	2.4	1lea	49	2.7	2.4
1xik.A	55	2.4	1avc	53	3.4	2.4
	⋮			⋮		

†The quantities RMSD and LALI in the DALI output correspond to L_1 and R_1, respectively (see Table III). DALI ranks similarity by a z-score which measures similarity with respect to a random background. The DALI data were downloaded from the FSSP data base.[11]

Multiple Structural Matches

As already mentioned the notion of a single optimum structural match or alignment is problematic. Multiplicity of solutions is a fundamental characteristic of protein structure similarity and must be taken into account explicitly. The range of possible rigid body matches embraces complete folds, domains, subdomains, secondary structure, and individual residues at the lowest level. The technical details used in ProSup to find multiple solutions were reported by Feng and Sippl.[8]

We define the term structural match to refer to a particular rigid body superimposition of two structures characterized by the set of equivalent residue pairs and the associated rms error. A match can be represented as an alignment, which in general contains gaps (Fig. 1). Matches can be compared and ranked, when rms values are constrained as discussed above.

For two folds A and B there is always a spectrum of matches $M_i(A, B)$, where $L_i(A, B)$ is the number of equivalent residues or alignment length of $M_i(A, B)$. The spec-

Fig. 1. Alternative alignments between transcriptional activator MarA (1bl0[10]) and diphtheria toxin repressor (2dtr[12]). M_1 and M_4 align the N-terminal domains of both proteins. Nevertheless, the alignments differ in the first and the third superimposed regions. In M_2 the C-terminal domain of 1bl0 is aligned with the N-terminal domain of 2dtr (see also Fig. 2). Shaded boxes and upper-case letters indicate structurally equivalent residues.

trum is ranked so that $L_i(A, B) \geq L_{i+1}(A, B)$ and we call $M_1(A, B)$ an optimum match or best alignment. The optimum can be degenerate, i.e., there may be several matches which have the same length. The optimum match is usually considered to represent the extent of structural similarity of two folds. For example, all the ProSup entries in Table I refer to $M_1(A, B)$.

In general however, the maximum match is insufficient for an exhaustive description of similarity among folds. To highlight the problem, we discuss two variants of alternative solutions using 1bl0 (target T0079, transcriptional activator MarA[10]) and 2dtr (diphtheria toxin repressor[12]) as an example. Both proteins are composed of two domains. The DNA binding domains in 1bl0 are superimposable representing a structural motif which matches the N-terminal DNA binding domain but not the dimerization domain in 2dtr.

The spectrum of structural similarity of 1bl0 and 2dtr shown in Table II contains 187 matches with $L_i \geq 20$. The most extensive match has $L_1 = 47$ equivalent residues. The spectrum varies rather continuously since there is no clear distinction between optimum and suboptimum alignments. Any two alignments in the spectrum of structural matches are distinct, but they may have several or many pairs of equivalent residues in common. The number of common pairs measures the similarity of two matches and is used to group the alignments in Table II. For example, the alignments M_2, M_3, and M_6 are closely related, whereas M_1, M_2, and M_4 are grossly distinct, although they may have some pairs in common as can be inferred from Figure 1.

The structural superimpositions corresponding to alignments M_1 and M_2 of Table II shown in Figures 1 and 2 match distinct domains, as might be expected from the internal repeat of 1bl0. On the other hand, M_1 and M_4 match the N-terminal domains but have quite distinct equivalent pairs (Fig. 1).

TABLE II. Alternative Alignments of 1bl0 and 2dtr

Number	L_i[a]	R_i[b]	Cluster[c]
1	47	2.7	3
2	45	2.7	1
3	44	2.7	1
4	43	1.8	10
5	43	2.1	3
6	43	2.3	1
7	43	2.3	1
8	43	2.3	3
9	43	2.3	3
10	43	2.6	1
11	43	2.6	1
12	42	2.1	10
13	42	2.1	10
14	42	2.5	1
15	41	2.2	1
		⋮	
186	20	3.0	66
187	20	3.0	67

[a]Number of equivalent residues of match M_i.
[b]Rms deviation of equivalent residues of match M_i.
[c]Cluster of match M_i. ProSup groups matches in a cluster which are more than 70% identical.

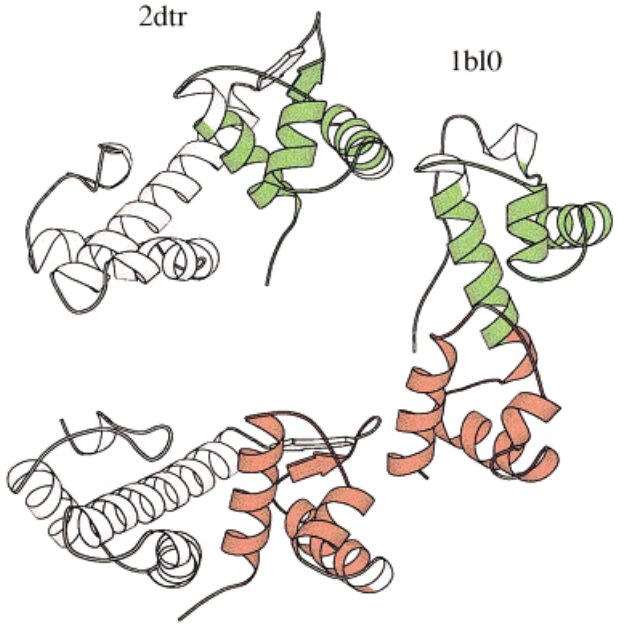

Fig. 2. Two possible superimpositions of transcriptional activator MarA (1bl0[10], right side) and diphtheria toxin repressor (2dtr,[12] left side, two different orientations). The two DNA-binding domains of MarA are structurally similar and both can be superimposed with the N-terminal DNA-binding domain of diphtheria toxin repressor. The alignments are not distinguishable geometrically. N-terminal domain (red): 47 eq. and 2.75 Å rms deviation. C-terminal domain (green): 45 eq. and 2.75 Å rms deviation. The C-terminal SH3-like domain of 2dtr was excluded for simplicity. The figure was prepared using Molscript.[23]

Evaluation of Predictions

There is no essential difference in the comparison of two structures A and B and the evaluation of a predicted structure P relative to a target T. In both cases the spectra represent the extent of structural similarity. The only distinction is that P and T are two conformations of a single protein and therefore, the residue positions relate to a common sequence. The equivalences implied by the sequence are independent from the equivalences implied by a structural match. Their relation can be used to measure model accuracy in terms of shifts.

To be precise, a structural match $M_i(P, T)$ between prediction P and target T is a sequence of pairs $(p_1t_1, p_2t_2, \ldots, P_{L_i}t_{L_i})$, where p_k and t_k refer to sequence positions of P and T, respectively. The difference $s(p_kt_k) = |p_k - t_k|$ is the shift of the aligned pair p_kt_k. The distribution of shifts in a match characterizes the accuracy of the respective alignment. The number of zero shifts of $M_i(P, T)$ counts the correctly aligned residues in the prediction. To evaluate the alignment accuracy of a prediction the match $M_x(P,T)$ has to be found which yields the maximum number of correctly aligned residues. Note that in general, the match with the largest number of equivalent residues *is not* the alignment with the maximum number of correctly aligned residues, $M_1(P,T) \neq M_x(P,T)$.

For example, assume that three predictors chose 2dtr as a template for T0079, where the predicted models correspond to the alignments M_1, M_2, and M_4 of Table II (models are generated by copying the sequence of 1bl0 onto the template 2dtr according to the alignments). Then an evaluation that takes into account only the optimum match M_1 will conclude that the prediction corresponding to M_1 is perfect, that the prediction corresponding to M_2 is completely wrong, and that the prediction corresponding to M_4 has a few correctly aligned residues.

Hence, we emphasize that an evaluation has to be based on the match having a maximum number of correctly aligned residues rather than the maximum match $M_1(P,T)$ (Fig. 3). Otherwise the evaluation might give rise to wrong conclusions.

Summary of Numbers

Table III summarizes the various numbers used in the CASP3 evaluation. $N(P,T)$ is the number of residues that can be compared between prediction P and target T. Usually this is the number of residues in a prediction, but it can be smaller when the target is incomplete. $R(P,T)$ is the root mean square error of superimposition of all N residues of P and T, sometimes called sequence dependent rms. $R(P,T)$ is a good measure when two structures are similar but it is not discriminative when structures have larger deviations. Therefore, it can be used to rank accurate predictions, but it is unsuitable in other cases. Another problem is, that $R(P,T)$ strongly depends on the lengths of P and T. It is often the case that one predictor may omit a difficult stretch of the structure, resulting in a lower $R(P,T)$ than that produced by another predictor who submits a complete model.

$L_1(P,T)$ and $R_1(P,T)$ correspond to the optimum structural match $M_1(P,T)$. L_1 is the number of aligned residues and R_1 the associated root mean square error. As discussed above R_1 is close to a constant value, which is close to 3 Å in the CASP3 evaluation.

The remaining numbers in Table III correspond to the structural match which yields the maximum number of correctly aligned residues for P, where L_x is the total length of this alignment. We did not include $R_x(P,T)$ since it is almost constant and in the same range as $R_1(P,T)$. $S0$ counts correctly aligned residues where $S0 = L_x$ corresponds to a perfect alignment. Shifts $S1$ to $S4$ indicate local alignment errors. $S5+$ summarizes large shift errors and finally the average shift SA is a measure for overall alignment quality.

The spectrum $M(A,B)$ always contains many local matches, like matches between secondary structure elements, which are generally considered to be insignificant. We therefore, truncate the spectrum for $L_i < 20$ and indicate this by zeros in the tables. Specifically, $L_x = 0$ means that there is no structural match which yields more than 19 correctly aligned residues. Finally, all rms calculations and superimposition of folds are based on C^α atoms.

CASP3 EVALUATION

Target structures and predicted models were obtained from the prediction server in PDB format. Sequence numbers and consistency of files were checked by the CASP prediction center.[13] Predictions were labeled by

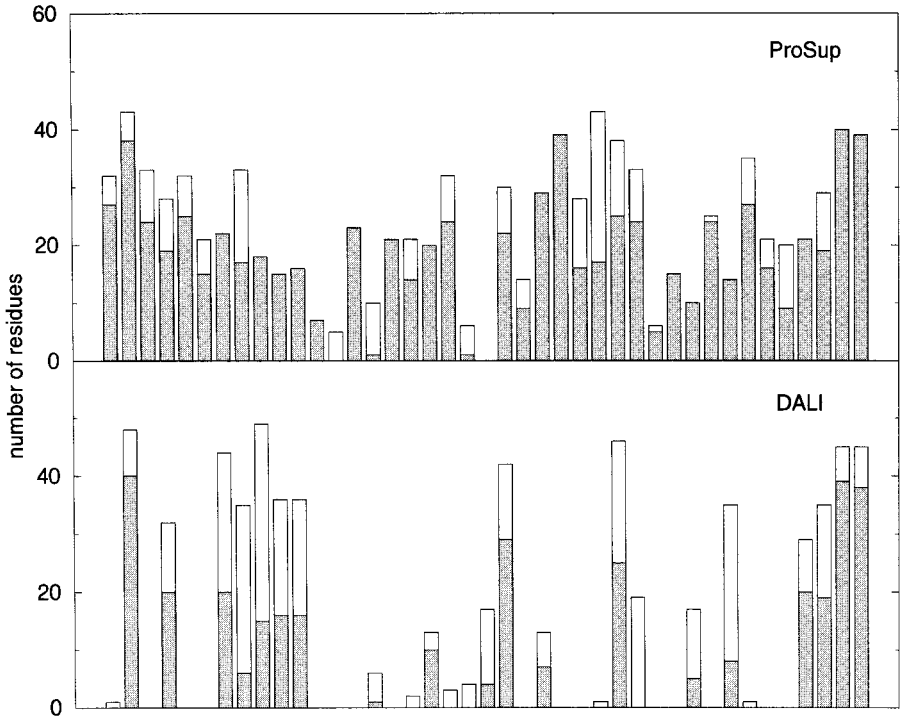

Fig. 3. Comparison of evaluation data for target T0079 models derived by ProSup (upper panel) and DALI (lower panel). Grey bars indicate the number of correctly aligned residues and the white bars the number of residues with a shift less than five. Both panels are sorted alphabetically by the name of the prediction group (bars in the same column refer to the one model). In many cases DALI superimposes model and target in an orientation that no residues are aligned correctly, although an alternative with correctly aligned residues exists. The DALI data from the CASP3 Server[13] were used for this comparison.

TABLE III. Numerical Evaluation Criteria

Measure	Synonym[a]	Description
N	nres	Total number of comparable residues
R	rms	Root mean square error of prediction and target
L_1	eqr1	Number of equivalent residues of optimum match M_1
R_1	rms1[b]	Root mean square error of optimum match M_1
L_x	eqrb[b]	Number of equivalent residues of match with the highest number of correctly aligned residues (M_x)
S_0	shf0	Count of correctly aligned residues (shift zero)
Sy	shfy[c]	Count of misaligned residues having shift $s = y$
$S5+$	shf5+	Count of misaligned residues having shift $s \geq 5$
SA	shf-av	Average over shifts

[a]Identifiers used in the tables at the CASP3 sever.[13]
[b]Not included in tables at the CASP3 server.[13]
[c]Shf1 to Shf4 are summed up in tables at the CASP3 server.[13]

submission number, so that predictors remained anonymous in the evaluation, except, of course, that we were able to recognize our own submissions. Targets T0045, T0050, T0051, T0062, T0069, T0072, and T0078 were not available for evaluation.

Specific questions in the evaluation of CASP experiments are:

- Was the fold novel or recognizable?
- How good are the predictions?

Below we demonstrate how these questions were addressed using target T0083 as an example. Evaluation data for other targets can be accessed at the CASP site.[13]

Was the Fold Recognizable?

CASP3 targets were compared to a representative library of 2,731 folds[14] derived from the September 9, 1998 data set of PDB.[15] The folds were ranked by $L_1(T,F_i)$, the length of the optimum match. Table IV shows the (truncated) ranked fold library for target T0083. There are two points to emphasize: (1) $R_1(T,F_i)$ varies only slightly so that $L_1(T,F_i)$ indicates the extent of similarity, and (2) there is a continuum of folds that are similar to the DNA binding domain of T0083.

How Good Are the Predictions?

This is a question of assessment rather than evaluation. Evaluation provides a numerical estimate but it is a matter of assessment to judge the significance of numbers and predictions. The goal here is to indicate how to interpret the numbers and how they can be used to support assessment. Table V assembles the results for predictions of T0083 as an example.

Table V ranks the predictions by L_1. This indicates the extent of structural similarity between prediction and target but does not take into account alignment accuracy. L_1 of Table V reveals that several predictions either used an appropriate template from PDB or somehow computed a fold that is similar to the best templates in PDB. The example also shows that there is no problem in comparing predictions across categories and that it is in fact advantageous to remove the boundaries.

TABLE IV. Database Search for T0083 With ProSup

PDB[a]	L_1[b]	R_1[b]	Release date[c]
pdb1lmb.3.-	64	2.3	Jan-31-1994
pdb1adr.-.1	60	2.4	Jan-31-1994
pdb1bab.A.-	58	2.9	Jan-31-1994
pdb2cro.-.-	57	2.3	Jan-31-1994
pdb1r69.-.-	57	2.7	Jan-31-1994
pdb1abw.A.-	57	2.8	Jun-17-1998
pdb1a4f.A.-	56	2.7	Apr-29-1998
pdb1neq.-.-	56	2.7	Dec-07-1995
pdb1spg.A.-	56	2.8	Mar-13-1997
pdb1out.A.-	55	2.8	Jan-13-1997
pdb1hds.B.-	53	2.6	Jan-31-1994
pdb1hda.B.-	53	2.6	Mar-23-1995
pdb1oct.C.-	53	2.6	Sep-15-1994
pdb1au7.A.-	53	2.7	Jan-28-1998
pdb1ak4.C.-	53	2.9	Oct-15-1997
pdb1waj.-.-	53	3.0	Jan-14-1998
pdb1apm.E.-	53	3.0	Jan-31-1994
pdb1ag8.A.-	52	2.9	Oct-08-1997
pdb1ann.-.-	52	2.9	Jan-31-1996
pdb1gpm.A.-	52	3.0	Jan-31-1996
⋮			

[a]PDB name including chain and model identifier (- if not available).
[b]See Table III.
[c]Time stamp of PDB file at the PDB FTP server.

$S0$ indicates alignment quality. Whenever $L_1 = L_x$ then the optimum structural match and the structural match yielding maximum $S0$ coincide. This is the case for the predictions T0083TS190_3 and T0083AL019_1, for example. For T0083AL017_2 L_1 and L_x are distinct but of the same order of magnitude. Hence the maximum match M_1 and the match M_x yielding the largest number of correctly aligned residues are of comparable size. In this specific case they also map to the same region in the molecule. In general the alignments may map to distinct regions even if L_1 and L_x are of comparable size (Fig. 1). In general, $L_x \ll L_1$ and/or $S0 \ll L_x$ indicate wrong or suboptimum alignments, as is the case in T0083TS005_2 and T0083AL040_1.

The two predictions, T0083AL028_1 and T0083AL176_1 have a comparatively small (conventional, sequence dependent) rms error of $R = 4.9$ and $R = 3.5$ Å, respectively. The alignment of T0083AL028_1 is perfect, whereas the alignment of T0083AL176_1 is somewhat suboptimal. In both cases the predictions are with respect to a domain rather than the whole sequence, as indicated by N. Other predictions contain more residues or predict the complete target. Although they have a comparable number of correctly aligned residues, parts of the predictions cannot be superimposed with the target fold giving rise to large R values and, therefore, they appear inferior as compared to T0083AL028_1 and T0083AL176_1.

It is difficult to draw a line between appropriate and inappropriate templates (Table IV) or between correct and incorrect predictions (Table V). Any such attempt is debatable. Hence, it is essential to have complete tables avail-

TABLE V. Evaluation Data for T0083

Model[a]	L_1[b]	R_1[b]	N[b]	R[b]	L_x[b]	$S0$[b]	$S1$[b]	$S2$[b]	$S3$[b]	$S4$[b]	$S5+$[b]	SA[b]	Template[c]
T0083TS190_3	66	2.6	156	14.0	66	66	0	0	0	0	0	0.0	n/a
T0083AL019_1	64	2.3	84	14.6	64	58	0	6	0	0	0	0.2	1lmb-3
T0083AL017_2	63	2.2	86	13.0	61	46	0	0	15	0	0	0.7	1lli-A
T0083AL028_2	62	2.2	81	12.9	60	49	7	4	0	0	0	0.2	1lli-A
T0083AL017_1	61	2.6	88	14.6	61	58	0	0	3	0	0	0.1	1lmb-4
T0083TS190_2	61	2.8	156	16.1	60	60	0	0	0	0	0	0.0	n/a
T0083TS005_1	61	2.2	84	10.0	59	59	0	0	0	0	0	0.0	1lmb-3
T0083AL028_1	60	2.4	65	4.9	60	60	0	0	0	0	0	0.0	1adr
T0083TS190_4	60	2.6	156	11.7	60	60	0	0	0	0	0	0.0	n/a
T0083AL176_1	57	2.7	63	3.5	55	49	6	0	0	0	0	0.1	1r69
T0083TS190_1	57	2.8	156	15.9	54	54	0	0	0	0	0	0.0	n/a
T0083TS005_2	53	2.6	117	17.6	23	23	0	0	0	0	0	0.0	1hbh-A
T0083AL040_1	50	2.6	137	17.5	29	15	6	8	0	0	0	0.8	1pbx-B
T0083AL061_1	44	3.0	143	17.5	44	0	0	5	12	0	27	20.3	2scp-A
T0083AL019_3	43	2.6	89	16.6	34	11	0	0	0	0	23	5.7	1ain
T0083TS035_4	43	2.8	156	16.1	37	37	0	0	0	0	0	0.0	n/a
T0083AL033_1	41	3.0	139	19.1	24	24	0	0	0	0	0	0.0	1rhg-A
T0083TS035_5	40	2.6	156	15.1	28	28	0	0	0	0	0	0.0	n/a
T0083TS061_1	40	2.8	139	17.1	40	0	0	0	0	0	40	30.2	2scp-A
⋮													
T0083AL215_1_1	0	100.0	16	4.6	0	0	0	0	0	0	0	0.0	1oro-B
T0083AL215_1_2	0	100.0	7	3.6	0	0	0	0	0	0	0	0.0	1oro-B

[a]Model identifier consisting of target number, prediction format (AL = alignment; TS = tertiary structure) and prediction group number.
[b]See Table III.
[c]Structural template identified by the predictor. In the case of ab initio predictions n/a is printed.

able for assessment. Most important numbers for ranking are $S0$, L_x and L_1, each of them highlighting particular aspects like sequence shifts and overall prediction quality. The program module CASPVIEW provides an interface for interactive analysis and supports ranking of prediction tables and is available on the CAME Web Site.[14]

CONCLUSION

The goal of this paper is to present the numerical criteria used in the evaluation and to provide a basis for their interpretation. At the start we had several goals and some remarks are in order for each of them. We refer to the items in the Introduction section.

First, the most important numbers in this evaluation are $S0$, L_x, and L_1. This is certainly a small set which nevertheless describes the most important aspects of a prediction. The remaining numbers highlight specific features, like local alignment shifts or overall root mean square error.

Methods for structure prediction are frequently categorized as comparative modeling, fold recognition or threading, and ab initio. They encompass a wide range of prediction techniques, but they have an obvious common goal: To compute a structural model that is as similar as possible to the experimental result. Hence, there is no fundamental distinction among prediction results. Therefore, the same general criteria can be used across categories (e.g., Table V). Of course, there are interesting details, like the quality of loops and side chain conformations in comparative modeling, deserving specific analysis.

Evaluation has to reveal the full extent of similarity between predictions and targets. Otherwise the analysis is inconclusive or even unfair. We paid particular attention to numerical problems and to an exhaustive enumeration of structural matches. The nonlinear nature of the structure comparison problem makes it impossible to guarantee that all optimal structural matches are found. But we can assure that it is rather unlikely that the CASP3 evaluation presented here missed important matches.

Frequently the comparison of protein structures results in multiple solutions. The analysis has to take into account the spectrum of solutions to find the best match between prediction and target. As we have discussed in some detail, prediction analysis is inconclusive if alternative solutions are neglected. These ambiguities have to be resolved in favor of the predictor.

The root mean square error is of comparable magnitude throughout the evaluation, so that the number of pairs of equivalent residues, or the length of an alignment measures the extent of a structural match. This removes at least one level of complexity as compared to scores that combine equivalences and root mean square error. Of course, ease of interpretation is a subjective issue and here we can only hope that at least the presentation is clear and transparent.

The main task of evaluation is to provide a numerical basis for assessment and no attempt is made here to judge the quality of predictions. Nevertheless, some comments are in order on the interpretation of numbers. In the previous section we pointed to the low rms values of T0083AL028_1 and T0083AL176_1. In other predictions the higher rms values are caused by regions outside the DNA binding domain. These regions are not predicted in T0083AL028_1 and T0083AL176_1 but within the domain the predictions are of comparable quality. The question arises how these differences should be assessed.

One standpoint is to reward T0083AL028_1 and T0083AL176_1 a bonus for confining the model to those parts that actually match the target fold. Another possibility for assessment is to focus on alignment quality, neglecting mismatches. We do not attempt to resolve this question here, but we emphasize that every predictor needs to know in advance how his prediction will be assessed. This comment applies to several other questions, e.g., how to measure the difficulty of targets, or how to take into account incorrect predictions. Such questions should be resolved as far as possible before the CASP4 prediction season starts.

The evaluation presented here is based on rigid body superimposition and therefore, has obvious limitations. There are other types of similarity that need different approaches. Examples are relative orientation of domains, distortions and twists. These are interesting and challenging problems which have been addressed in the past,[16–21] but they need careful analysis before they are applied to large scale evaluations.

There are many other aspects of evaluation that are potentially interesting. For example, alignments could be evaluated in more detail by using some measure of alignment compactness or the match of active site residues could be determined. Also, some questions are more relevant for one prediction category than another, like side chain configuration and loop quality in comparative modeling. But it is clear that each additional item adds to the complexity of evaluation and assessment.

In highlighting problems of structure comparison we frequently referred to DALI. This should not be seen as a criticism. In fact DALI was not designed for prediction evaluation but it is intended to provide a valuable service to the protein structure community. Regarding ProSup, the exercise to implement this evaluation scheme resulted in many insights and improvements that would not have been possible to obtain otherwise. Some of these are fundamental issues for structure comparison. Information about the availability of ProSup can be found on the internet.[22]

ACKNOWLEDGMENTS

We thank all X-ray crystallographers and NMR spectroscopist for providing prediction targets. We are also grateful to J. Moult, T. Hubbard, K. Fidelis, and the Lawrence Livermore Prediction Center team for providing the data and for helpful suggestions. F.S.D. acknowledges a fellowship of the Fundação para a Ciência e a Tecnologia (grant PRAXIS XXI/BD/4528/94).

REFERENCES

1. Mosimann S, Meleshko R, James MNG. A critical assessment of comparative molecular modelling of tertiary structures of proteins. Proteins 1995;23:301–317.
2. Lemer CM, Rooman MJ, Wodak SJ. Protein structure prediction by threading methods: evaluation of current techniques. Proteins 1995;23:377–355.
3. Defay T, Cohen FE. Evaluation of current techniques for ab initio protein structure prediction. Proteins 1995;23:431–445.
4. Marchler-Bauer A, Bryant SH. Measures of threading specificity and accuracy. Proteins Suppl 1997;1:74–82.
5. Venclovas C, Zemla A, Fidelis K, Moult J. Criteria for evaluating protein structures derived from comparative modeling. Proteins Suppl 1997;1:7–13.
6. Zemla A, Venclovas C, Reinhardt A, Fidelis K, Hubbard TJ. Numerical criteria for the evaluation of ab initio predictions of protein structure. Proteins Suppl 1997;1:140–150.
7. FEBS Advanced Course, Frontiers of protein structure prediction; 1997. http://predict.sanger.ac.uk/irbm-course97/.
8. Feng ZK, Sippl MJ. Optimum superimposition of protein structures: ambiguities and implications. Fold Des 1996;1:123–132.
9. Lackner P, Koppensteiner WA, Domingues FS, Sippl MJ. Recurrent problems in protein structure comparison. Submitted, 1999.
10. Rhee S, Martin RG, Rosner JL, Davies DR. A novel DNA-binding motif in MarA: the first structure for an AraC family transcriptional activator. Proc Natl Acad Sci USA 1998;95:10413–10418.
11. Holm LL, Sander C. Mapping the protein universe. Science 1996;273:595–602.
12. Qiu X, Pohl E, Holmes RK, Hol WG. High-resolution structure of the diphtheria toxin repressor complexed with cobalt and manganese reveals an SH3-like third domain and suggests a possible role of phosphate as co-corepressor. Biochemistry 1996;35:12229–12302.
13. CASP3 Home Page. http://PredictionCenter.llnl.gov/casp3/Casp3.html.
14. CASP Evaluation Page at CAME. http://www.came.sbg.ac.at/CASP/.
15. Sussman JL, Lin D, Jiang J, Manning NO, Prilusky J, Ritter O, Abola EE. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. Acta Crystallogr D 1998;54:1078–1084.
16. Cohen F, Sternberg MJ. On the prediction of protein structure: the significance of the root-mean-square deviation. J Mol Biol 1980;138:321–333.
17. Sippl MJ. On the problem of comparing protein structures. Development and applications of a new method for the assessment of structural similarities of polypeptide conformations. J Mol Biol 1982;156:359–388.
18. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
19. Orengo CA, Taylor WR. SSAP: sequential structure alignment program for protein structure comparison. Methods Enzymol 1996;266:617–635.
20. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6:377–385.
21. Lesk A. Extraction of geometrically similar substructures: least-squares and Chebyshev fitting and the difference distance matrix. Proteins 1998;15:320–328.
22. ProCeryon—A software package for fold recognition and protein structure analysis, King's Beech Biosoftware, http://www.kings-beech.com, 1999.
23. Kraulis PJ. Molscript: a program to produce both detailed and schematic plots of protein structures. J Appl Cryst 1991;24:946–950.