PROTEINS: Structure, Function, and Genetics Suppl 3:15–21 (1999)

RMS/Coverage Graphs: A Qualitative Method for Comparing Three-Dimensional Protein Structure Predictions

Tim J.P. Hubbard*
Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, United Kingdom

ABSTRACT **Evaluating a set of protein structure predictions is difficult as each prediction may omit different residues and different parts of the structure may have different accuracies. A method is described that captures the best results from a large number of alternative sequence-dependent structural superpositions between a prediction and the experimental structure and represents them as a single line on a graph. Applied to CASP2 and CASP3 data the best predictions stand out visually in most cases, as judged by manual inspection. The results from this method applied to CASP data are available from the URLs http://PredictionCenter.llnl.gov/casp3/results/th/ and http://www.sanger.ac.uk/~th/casp/. Proteins Suppl 1999;3:15–21.**

Key words: protein structure prediction; protein structure comparison; CASP

INTRODUCTION

The most common way to compare pairs of three-dimensional protein structures has long been root mean square distance (RMS) superposition. This has been used in various ways in the evaluation of predictions in previous CASP experiments.[1–9] Many algorithms have been developed to find the optimal superposition, mostly focusing on the difficulty of obtaining the correct set of structurally equivalent residues (the structural alignment) when two different structures are being compared. In the special case of comparing a predicted structure with an experimental result, this alignment is known (although it is still useful to explore alternative alignments,[10] see also Sippl et al.[13]) since the two sequences are identical, which reduces the complexity of the problem. On the other hand, the problem is no longer necessarily the comparison of two objects on the basis of there being a real structural relationship to identify. Bad predictions may seem almost totally dissimilar to the target structure, but still have to be compared to identify if there is anything that has been predicted usefully. Methods that can be applied to evaluation of predictions therefore have to be robust to very large variations in structural similarity.

In the case of evaluating many predictions, what is also required is a way of comparing the different results in an objective way. The problem with this is that no two predictions are the same. Different groups may predict different subsets of the residues in the structure and may predict different parts of the structure at different accuracies. How can we evaluate what is the better prediction, when one is the entire structure, and another is just the core of the protein? Any cutoff that is applied, on the basis of RMS or number of equivalent residues, is likely to disadvantage some predictions, however without any cutoff it would seem that we are likely to drown in a sea of numbers.

One approach to this problem is to present the results of a very large number of superpositions graphically. A large number of superpositions are used to sample the best RMS for each number of equivalent residues (not necessarily contiguous). The graphical representation is a line for each prediction relating these best RMS values to number of equivalent residues. The result is the RMS/Coverage graph, which appears to represent the best prediction as the lowest line on the graph for most CASP2 and CASP3 targets, as judged by the manual inspection of the assessors. The results can be viewed at http://PredictionCenter.llnl.gov/casp3/results/th/ for CASP3 data (some 3D coordinates are not available yet as they are still unpublished), and from http://www.sanger.ac.uk/~th/casp/ for CASP2 data.

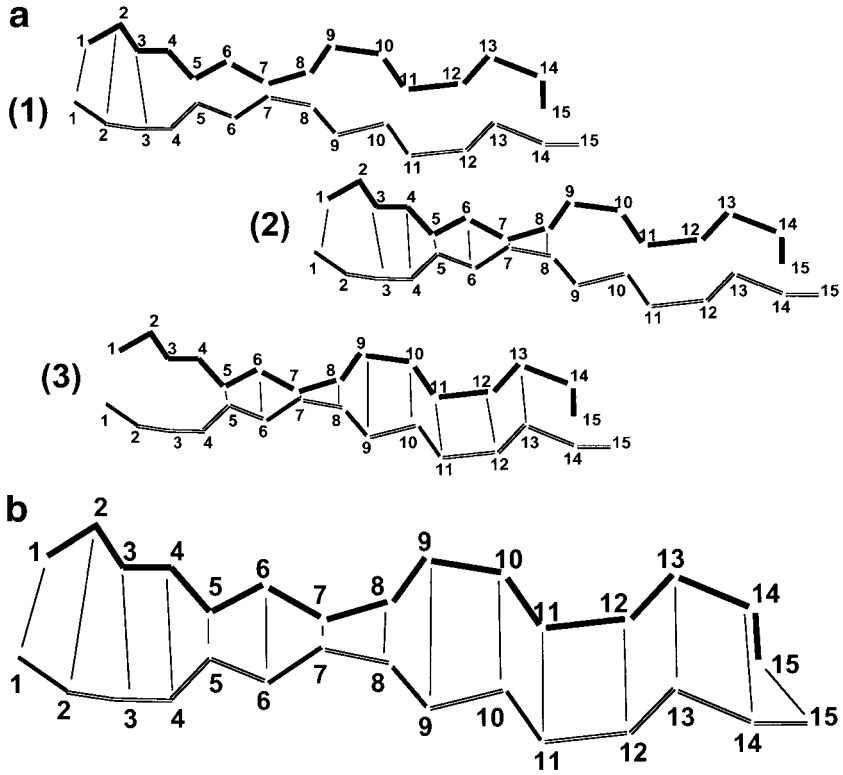
MATERIALS AND METHODS

Data

Predictions in CASP2 were submitted in three different formats: CM (comparative modeling), FR (fold recognition/threading) and AB (ab initio). After excluding AB predictions at 1D (secondary structure) and 2D (contact) levels, all remaining predictions are or can be represented as a 3D structure. In the case of AB and CM, the remaining predictions were submitted as coordinate sets. In the case of FR, predictions were submitted as alignments, so 3D models were created from the first five alignments to which the predictor gave the most confidence. Models were created using the program AL2TS[11] by Adam Zemla at the Prediction Centre.

For CASP3 data there is no separate CM/FR/AB submission, however there are still different data formats of SS (secondary structure), RR (residue separation), AL (alignment to known structure) and TS (tertiary structure). TS submissions are already 3D coordinate sets and for AL submissions 3D models were again created from the alignment using AL2TS.

*Correspondence to: Tim J.P. Hubbard, Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, United Kingdom. E-mail: th@sanger.ac.uk
Received 4 June 1999; Accepted 14 June 1999



(1) ranked list of pairs

7 - 7
5 - 5
8 - 8
15 - 15
6 - 6
1 - 1
11 - 11

(2) rms coverage list

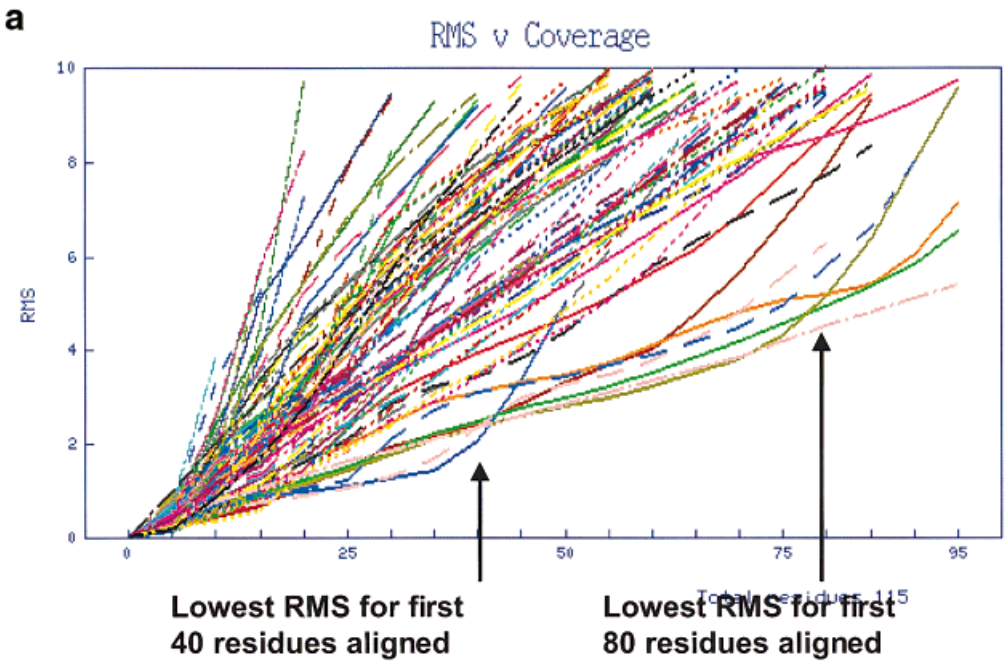
1	7	1.2A
2	7,5	1.5A
3	5,7,8	1.8A
4	5,7,8,15	1.9A
5	5-8,15	2.0A
6	1,5-8,15	2.2A
7	1,5-8,11,15	2.5A

Fig. 1. **a:** A schematic representation of iterative superposition. (1) The initial superposition is made using the coordinate pairs of residues 1–3 in the two structures. (2) After the first superposition residue pairs 1,3,4,5,6,7 and 8 are closer than the cutoff of six Angstroms and are used to make the next superposition. (3) After the second superposition the position of the structures has shifted such that residue pairs 5–13 are closer than the cutoff. These would be used to make the third superposition. In most cases, continuing this process will result in convergence to a list that no longer changes. **b:** A schematic representation of the construction of an RMS coverage list from a superposition. The distances between all equivalent pairs of residues are calculated. These distances are sorted smallest to largest to give (1). The RMS is calculated from the distances for a coverage of 1 (top pair); 2 (first two pairs); 3 (first three pairs) etc., to give the values in (2).

Generating a Large Number of Structural Superpositions

A structural superposition results from the unique transformation, which minimizes the RMS between two lists of atomic coordinates. Different superpositions therefore result from different lists. In this algorithm the lists are generated by iterating from all possible starting points of three consecutive results. For example, for a protein of 100 residues the starting points are residues 1–3 of the prediction superposed with residues 1–3 of the target; residues 2–4 superposed with residues 2–4; residues 3–5 etc., up to residues 98–100 (see Fig. 1a).

Iteration consists of building a new list from the result of the previous superposition, followed by a new superposition, etc. The new list is constructed by measuring the distance between equivalent residues. Any pair for which the distance is less than six Angstroms is included in the new list. Six Angstroms is a high distance threshold compared with that used normally (~3 to 3.5 Angstroms) in structure comparison methods. It can afford to be so high because sequence-dependent structural superposition is being carried out here so what is an equivalent residue is unambiguous. It is useful to make this high, because of the RMS between a predicted structure and target can be much larger than between two naturally similar protein structures as predictions do not necessarily respect protein geometry. In this experiment, three iterations are carried from each starting point. For many prediction/target pairs, after three iterations many of these superpositions and their corresponding residue pair



b

Analysis of Casp3 predictions: T0046: RMS 0-10 ...

[1st prediction | All predictions]

T0046 (115)

[Full range of RMS | RMS 0-10 A | RMS 0-10 A, first 100 residues | Full range of RMS, colour by group]

RMS v Coverage

RMS

Total residues 115

T0046AL074:	STERNBERG	1	106	**3.56**	[p] [i] [s] 3hhrc:2.1.2 [24/10]
T0046TS061:	YANG-ANSUEI	1	109	**3.65**	[p] [i] [s] 3hhrc:2.1.2 [24/10]
T0046AL061:	YANG-ANSUEI	1	98	**3.98**	[p] [i] [s] 3hhrc:2.1.2 [24/10]
T0046AL212:	BRYANT	1	84	4.35	[p] [i] [s] 3dpa_:2.1.8 [24/2] ,2.6.2 [1/1]
		2	90	**4.24**	[p] [i] [s] 1mspa:2.1.8 [24/2]
		3	79	9.47	[p] [i] [s] 1jrhi:2.1.2 [24/10]
		4	86	6.41	[p] [i] [s] 3dpa_:2.1.8 [24/2] ,2.6.2 [1/1]
		5	68	14.74	[p] [i] [s] 1jrhi:2.1.2 [24/10]
T0046AL017:	SJOLANDER-THOMAS	1	73	11.69	[p] [i] [s] 2mcm_:2.1.6 [24/4]
		2	102	**4.56**	[p] [i] [s] 1ggta:2.1.1 [24/12] ,4.3.1 [1/1] ,2.1.4 [24/1]
T0046AL003:	HUBBARD	1	86	**4.96**	[p] [i] [s] 1fna_:2.1.2 [24/10]
		2	104	6.25	[p] [i] [s] 2mcm_:2.1.6 [24/4]
T0046TS035:	BAKER	1	119	**7.03**	[p] [i] [s] N/A:

Fig. 2. RMS/coverage graph for T0046 is shown with axes of 1–100 residues coverage and 0–10 Angstroms RMS (see text for discussion). **(a)** shows graph itself, with arrows indicating key features. **(b)** shows the graph as it appears on a web page with its associated text.

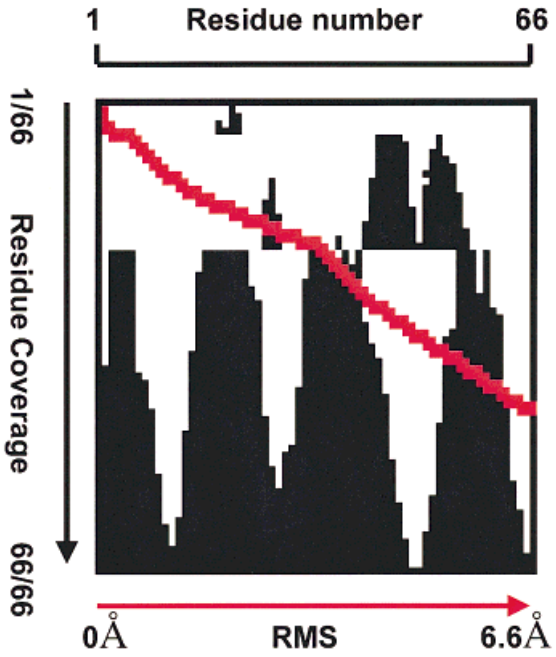


Figure 3.

Fig. 3. Equivalent residue/Coverage graph for an ab initio CASP2 prediction of target T0030 (T0030AB807). The red line shows the RMS/Coverage line for this prediction taken from the graph for all predictions for this target. Plotted on the same graph, with the same X axis, is a dot plot where the Y axis is the residue numbering of the target protein. Each black pixel indicates that the single residue is part of the equivalent list making up the coverage on the X axis (see text for discussion).

Fig. 4. Two predictions of the B domain of T0063 are shown. The predictions are T0063bTS005 **(a)** and **(b)** and T0063bAL066 **(c)** and **(d)** (prediction T0063bTS005 is a 3D model submission by group 5; T0063bAL066 is an alignment submission by group 66). The coloring schemes are based on distance separation for (a) and (c): $<$ 2A: yellow (residue pairs not shifted and well modeled); 2–4A: orange (badly modeled/shifted by 1 residue); 4–8A: red (badly modeled/shifted by 1–2 residues); 8–12A: violet (badly modeled/shifted by 3–4 residues); $>$ 12A: purple and segment shift after distance constrained Smith-Waterman alignment for (b) and (d): no shift: yellow; shift of 1–2: orange; shift of 3–5: red; shift of 5–10: violet; shift of $>$ 10: purple. With a suitably configured browser/helper the (a) and (c) structures can be viewed by clicking on the link labeled [i] on the web pages and the (b) and (d) predictions viewed by clicking on the link labeled [s].

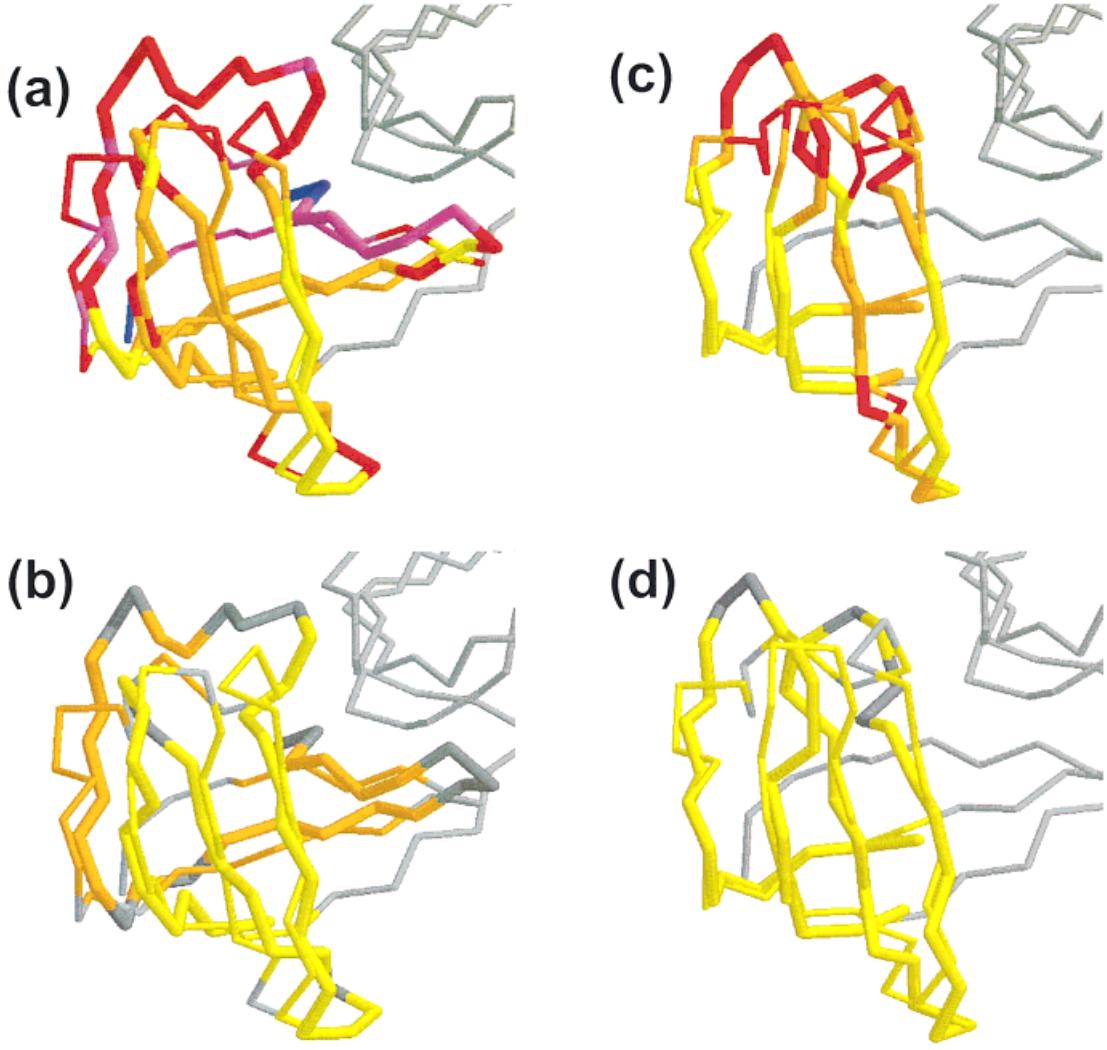


Figure 4.

lists will be very similar, i.e., the iteration converges. In cases where there are several quite dissimilar end points the prediction contains different substructures that are similar to the target but which cannot all be superposed together. For the evaluation carried out here the 98*4 different superpositions for a 100 residue protein appears to sample superposition space sufficiently for the next step.

Determining the Minimum RMS for Each Coverage Value

Coverage is defined here as the fraction of the target being predicted for the number of residues being considered. For a 100 residue protein coverage values range from 0.01 (1/100 residues) to 1 (100/100 residues). Coverage as defined here is non-consecutive, i.e., residues 1,3,10,88 can be considered as a coverage of 0.04 just as residues 1–4 can. The minimum RMS for each coverage value out of all the superpositions sampled can be determined by measuring the distance between each equivalent residue pair; sorting this list and then calculating the RMS for the first two residue pairs in the list, the first three pairs, etc. (see Fig. 1b). The minimum RMS for each coverage value for the entire prediction/target comparison is the set of lowest RMSs for each coverage value, across all the different superpositions. It is this minimum RMS that is plotted against coverage as a line on an RMS/Coverage graph (see Fig. 2).

RESULTS

RMS/Coverage Graphs

An example of a complete RMS/Coverage graph is shown in Figure 2 for CASP target T0046. This example illustrates how such graphs allow several good predictions with different features to be picked out from the rest. It can be seen that there are a couple of predictions with very low RMS at about 40 residues coverage, but whose RMS rapidly rises after that, and some more predictions with worse RMS at 40 residues, but the best RMS for much higher coverage values (80 residues). The annotated graphical displays of these two predictions cannot be printed here, however, what they would show is that the 40 residue prediction is of the core secondary structures of the protein (a set of beta strands), but does not include loops. The 80-residue prediction is of the full structure, but obviously has a core that is not as accurate as the 40 residue prediction. In the context of the CASP experiment, the graph draws attention to interesting predictions that should be examined in more detail.

Equivalent Residue/Coverage Plots or Each Prediction

Since each line on an RMS/Coverage graph can be made up of results from many different superpositions, it is useful to know which residue pairs were included at any given coverage value. Equivalent residue/Coverage plots are therefore calculated for each prediction. Figure 3 shows the equivalent residue/Coverage plot for the CASP2 prediction T0030AB807. On all plots, one pixel corresponds to one residue (black) or 0.1A RMS (red line, Y axis). The exact correspondence between residue and RMS is to ensure that the slope of the red line is comparable between plots. T0030 has a length of 66 residues, so the point at which the red line touches the maximum on the Y axis is when the RMS is 6.6A. The black dot plot shows the different superpositions that give rise to the lowest RMS for each coverage point. Working along the X axis, the first block of black pixels can be seen at residues 20–23. This corresponds to the best superpositions sampled involving two, three, or four residues. For five residues, this region no longer has the best superposition. It is superseded by a region centered on the hairpin around residues 48–49. This region is built up until 18 residues are involved in the superposition, mostly involving this single well predicted hairpin (superposition of these 18 residues is ~3A). To superpose 19 residues, however, a completely different set of equivalences gives the lowest RMS, involving parts of the entire prediction. What this shows, without needing to look at the structures, is that regions throughout the prediction can be superposed at around 4A RMS, but that there is a piece of structure centered on residue 48, which is predicted correctly locally, but not with respect to the rest of the structure.

Selection of the Best Superposition and Method for Graphical Display

While sampling the different superpositions, a record is kept of the one that results in the largest list of equivalent pairs. The coordinates of this superposition are kept and are linked to the RMS/Coverage graph page of the website. As well as linking raw coordinates, two annotated coordinate sets are provided in RasMol inline format. The annotation is to color code residue pairs according to their accuracy.

In the two alternative coordinate sets, two annotation methods are used (see Fig. 4). In the first, residue pairs are colored according to their distance separation. The coloring scheme used is: $<$ 2A: yellow (residue pairs not shifted and well modeled); 2–4A: orange (badly modeled/shifted by 1 residue); 4–8A: red (badly modeled/shifted by 1–2 residues); 8–12A: violet (badly modeled/shifted by 3–4 residues); $>$ 12A: purple.

In the second, residue pairs are colored according to the shift between the (correct) sequence based alignment and the alignment implied by the structural superposition, i.e., if residues 3 and 6 are structurally aligned, the shift is 3. The alignment implied by the structural superposition is calculated by a Smith-Waterman algorithm applied to the full distance matrix resulting from the superposition implemented using the dynamite dynamic programming compiler.[12] Although the first coloring scheme relates distances to a rough number of shifts, the effect of the Smith-Waterman algorithm is to smooth the output so that entire secondary structures are colored according to their alignment shift, rather than each residue being different. The coloring scheme used is: no shift: yellow; shift of 1–2: orange; shift of 3–5: red; shift of 5–10: violet; shift of $>$ 10: purple.

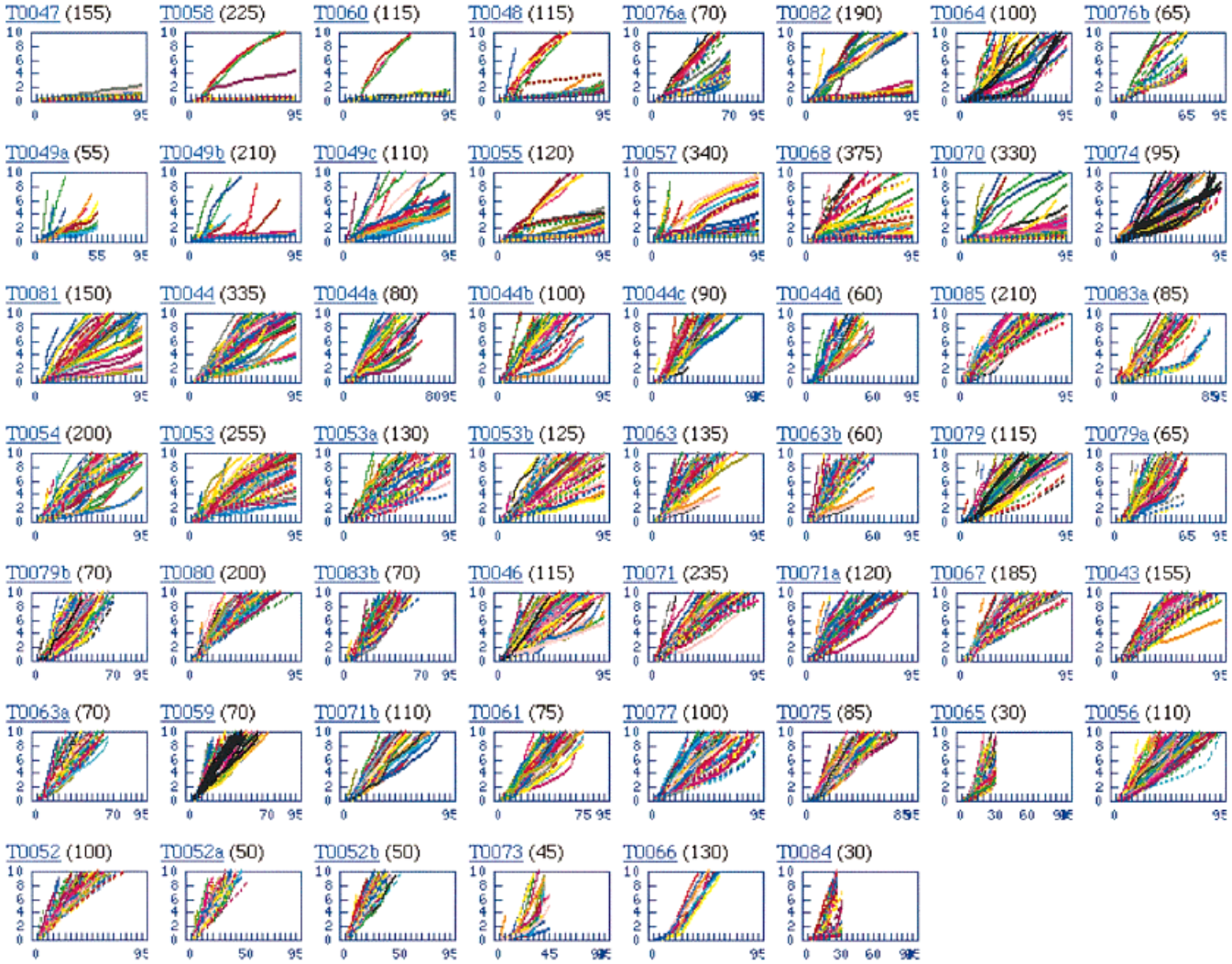


Fig. 5. Thumbnail images of the RMS/Coverage graphs for CASP3, roughly ordered by difficulty.

Two predictions for the b domain of T0063 can see seen annotated in this way in Figure 4. The predictions are T0063bTS005 (a) and (b), and T0063bAL066 (c) and (d). In (a) and (c) every residue in the prediction (thick lines) is colored, however it is hard to work out exactly what the problems in superposition of prediction and target are. In (b) and (d), some residues between segments of the prediction are not aligned to the target and are colored gray, however, it is much clearer that whereas T0063bAL066 has been superposed with no overall segment shifts, the superposition of T0063bTS005 leads to a one or two residue shift in one face of the beta barrel. Note that these views do not evaluate a prediction, but are ways of simplifying the viewing of a particular structural superposition.

Website

All of the above data is available for CASP3 predictions from the address http://PredictionCenter.llnl.gov/casp3/results/th/. The data is presented in a number of different ways to make it easy to view predictions from the point of view of a particular predictor or a particular target. For example, when making RMS/Coverage graphs there is an issue of what axes to use. Constant units for both RMS (y axis) and Coverage (x axis) mean that the slope of the curve can be compared between predictions and indicates their relative difficulties. Graphs are therefore provided showing the first 100 residues coverage against 0–10A RMS. However, this masks good performance on very large predictions and does not show the full RMS range, so two more sets of graphs are provided: 1) the full coverage scaled onto the x axis with 0–10A RMS range and 2) full coverage and full RMS range.

In CASP3, predictors were allowed to submit up to five predictions for each target, but it was emphasized that the assessment would concentrate on their first prediction. As a result, there are two graphs for each target, the first just showing the top predictions and the second showing all five predictions.

Figure 2 shows the 'All predictions' graph for T0046. Figure 2b shows the graph with the text as it appears on the web page. Underneath each graph, information is listed with one line for each prediction. The information shown is: (1) prediction ID; (2) predictor group; (3) number of predictions; (4) the number of residues predicted; (5) an RMS; (6) links labeled [p], [i], [s]; (7) the identifier of the PDB chain from which the model is predicted; (8) the SCOP fold code for the PDB code; and (9) the frequency with which this code is used over all predictions for this target. If the prediction is ab initio, 7,8,9 are undefined. The RMS given is for a coverage of 66 residues and the list of predictions are ordered according to this value. When the graphs show all predictions made, the best RMS by each predictor is shown in bold and the ranking is according to this value. In the case where fewer than 66 residues were predicted, the RMS value is an extrapolation and shown in italics. On such pages, holding the mouse over any of the lines on the graph causes the name of that prediction to appear in the status bar of the browser, and clicking will jump down the page to the line referring to that prediction. (These images are drawn using Perl Modules GD.pm and GIFgraph.pm, extended by Matt Pocock available from the common perl archive network CPAN). In the text, clicking on the prediction number (3) will jump to the 'Equivalent residue/Coverage graph' for that prediction. With a rasmol viewer correctly configured as a helper application of your browser, clicking on [p], [i] or [s] will cause the Calpha coordinates (with residue coloring annotation in case of [i] or [s]) to be downloaded and displayed.

As well as these 'per prediction' pages, there are pages showing thumbnail images of the RMS/Coverage graphs by target and by group. Figure 5 shows the page with images of all predictions roughly ordered by difficulty. These images show the first 100 residues coverage against 0–10A RMS, so the slope is constant. It is immediately apparent the relative quality of the predictions: for T0047 all predictions are clearly accurate homology models, since the curve is flat, however for T0058 it can be seen that a few groups made wrong predictions. It becomes apparent what part of the graph approximates to a random prediction and it can be seen where there are lines that fall below this and are better.

In the title, the word 'qualitative' is used to describe the method. This is to emphasis that the method does not generate a single universal number that allows predictions to be ranked, but rather a graphical view that is clear to the eye. The ranking of predictions on the web pages (Fig. 2b) is based on a single number extracted from data used to generate the plot, but is only one possible ranking. The parameter used (lowest RMS for a coverage of 66 residues) seems a good compromise for many predictions, but does not reproduce what the eye would identify as the best in all cases, which is a limitation of the method. Figure 5 shows how the method allows many predictions to be compared in a visual way. However, since no satisfactory way has yet been found to convert the data in these images into a single number, the method has not been used to quantify performance on different targets.

CONCLUSIONS

Making sense of the vast amount of CASP data is very difficult and takes a great deal of time. Automatically generated analysis presented in a graphical 'point and click' way helps assessors and predictors focus on the most interesting predictions, by showing at a glance the range of predictions for any target. These interfaces to CASP analysis data will continue to be developed and extended to integrate all historical CASP data into a single view and extend the range of ways in which it is analyzed.

ACKNOWLEDGMENTS

I am very grateful to Anna Tramontano, John Moult, Alexey Murzin, Christine Orengo, Manfred Sippl, Alfonso Valencia, and others for many useful comments and discussions concerning these graphs and CASP prediction assessment in general. I am grateful for help from Adam Zemla for generating the coordinates from alignment submissions for CASP2 and CASP3, Ewan Birney in the use of dynamite to generate alignments from distance-distance matrices, and Matt Pocock for extending the GIFgraph perl modules used to generate the website.

REFERENCES

1. Mosimann S, Meleshko R, James MN. A critical assessment of comparative molecular modeling of tertiary structures of proteins. Proteins 1995;23:301–317.
2. Lemer CMR, Rooman MJ, Wodak SJ. Protein structure prediction by threading methods: evaluation of current techniques. Proteins 23:337–355.
3. Defay T, Cohen FE. Evaluation of current techniques for ab initio protein structure prediction. Proteins 23:431–445.
4. Martin AC, MacArthur MW, Thornton JM. Assessment of comparative modeling in CASP2. Proteins Suppl 1997;1:14–28.
5. Venclovas Č, Zemla A, Fidelis K, Moult J. Criteria for evaluating protein structures derived from comparative modeling. Proteins Suppl 1997;1:7–13.
6. Levitt M. Competitive assessment of protein fold recognition and alignment accuracy. Proteins Suppl 1997;1:92–104.
7. Marchler-Bauer A, Bryant SH. Measures of threading specificity and accuracy. Proteins Suppl 1997;1:74–82.
8. Lesk AM. CASP2: report on ab initio predictions. Proteins Suppl 1997;1:151–166.
9. Zemla A, Venclovas Č, Reinhardt A, Fidelis K, Hubbard TJP. Numerical criteria for the evaluation of ab initio predictions of protein structure. Proteins Suppl 1997;1:140–150.
10. Feng ZK, Sippl MJ. Optimum superimposition of protein structures: ambiguities and implications. Fold Des 1996;1:123–132.
11. Zemla A, Venclovas Č, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins Suppl 1999;3: 22–29.
12. Birney E, Durbin R. Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. Ismb 1997;5:56–64.
13. Sippl MJ, Lackner P, Domingues FS, Koppensteiner WA. An attempt to analyse progress in fold recognition from CASP1 to CASP3. Proteins Suppl 1999;3:226–230.