PROTEINS: Structure, Function, and Genetics Suppl 5:39–46 (2001)
DOI 10.1002/prot.1168

Enhancement of Protein Modeling by Human Intervention in Applying the Automatic Programs 3D-JIGSAW and 3D-PSSM

Paul A. Bates,* Lawrence A. Kelley, Robert M. MacCallum, and Michael J.E. Sternberg*
Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, London, United Kingdom

ABSTRACT **Fourteen models were constructed and analyzed for the comparative modeling section of Critical Assessment of Techniques for Protein Structure Prediction (CASP4). Sequence identity between each target and the best possible parent(s) ranged between 55 and 13%, and the root-mean-square deviation between model and target was from 0.8 to 17.9 Å. In the fold recognition section, 10 of the 11 remote homologues were recognized. The modeling protocols are a combination of automated computer algorithms, 3D-JIGSAW (for comparative modeling) and 3D-PSSM (for fold recognition), with human intervention at certain critical stages. In particular, intervention is required to check superfamily assignment, best possible parents from which to model, sequence alignments to those parents and take-off regions for modeling variable regions. There now is a convergence of algorithms for comparative modeling and fold recognition, particularly in the region of remote homology. Proteins 2001;Suppl 5:39–46.** © 2002 Wiley-Liss, Inc.

Key words: CASP4; bioinformatics; protein prediction; comparative protein modeling; homology modeling; protein fold recognition

INTRODUCTION

There is a major demand for fully automated approaches for the prediction of protein three-dimensional structure.[1] This demand has sharpened over recent years because of the large number of complete genome sequences now available (see, e.g., The Institute for Genomic Research (TIGR) web site at: http://www.tigr.org). However, results from previous Critical Assessment of Techniques for Protein Structure Prediction (CASP) suggested that fully automated prediction procedures are less accurate than approaches using some human intervention.[2,3] Why is this so? Can automated methods be improved to match the added value humans input to structure prediction protocols? These questions can now start to be addressed. In particular, those laboratories with fully automated servers as well as expertise at protein structure prediction are in a good position to address such questions. Here we describe our automatic methods, 3D-JIGSAW designed for comparative modeling and 3D-PSSM developed for fold recognition, and how the results from these programs were enhanced by our expertise in protein structure.

Within the laboratory, the two programs were designed, written, and developed separately. We thought this was prudent because the extent to which comparative modeling and fold recognition overlap was not clear. However, it was always the intention to compare the two approaches, and CASP4 provided the ideal time for this evaluation.

MATERIALS AND METHODS

CASP4 and CAFASP2 Submissions

Both prediction programs 3D-JIGSAW (comparative modeling) and 3D-PSSM (fold recognition) were entered into CASP4 and CAFASP2. The CASP experiments allow predictors to submit models predicted by any methodology, which may include human intervention. In contrast, the CAFASP experiments require that models are generated and submitted entirely automatically.

The program 3D-JIGSAW was originally designed to work on targets found to have at least one parent structure with no less than 40% sequence identity between target and parent(s). It was envisaged that there would be many such examples for CASP4/CAFASP2 (see article by Fischer et al. for definition and results to CAFASP2). Thus, it was thought that a very fine alignment algorithm might be needed. However, close monitoring of the CAFASP2 web site during the summer of 2000 indicated that there were to be very few targets with high sequence similarity to a parent homologue. Thus, we decided to use a more fold recognition-based alignment algorithm for our submissions to CASP4; these alignments were checked and, in some cases, manually adjusted. In addition, after the completion of CAFASP2 (October 2000) but before any of the comparative modeling targets were published or recorded in structural databases, we ran 3D-JIGSAW in full automatic mode on the comparative modeling targets again. This time we used the new and current alignment algorithm (see below for description of the two alignment algorithms).

Paul Bates and Lawrence Kelley contributed equally to this work.

Robert M. MacCallum's present address is Stockholm Bioinformatics Center, Department of Biochemistry, Stockholm University, S-106 91 Stockholm, Sweden.

*Correspondence to: Paul A. Bates or Michael J.E. Sternberg, Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, London WC2A 3PX, UK. E-mail: paul.bates@ICRF.icnet.uk or m.sternberg@ICRF.icnet.uk

Received 12 March 2001; Accepted 12 June 2001

Methods for 3D-JIGSAW[4] and 3D-PSSM[5] have previously been reported; thus, only a brief outline will be given for each program.

Sequence and Protein Coordinate Databases Used by 3D-JIGSAW and 3D-PSSM

Each week our local copy of the nonredundant protein sequence database (nr) (ftp://ncbi.nlm.nih.gov/blast/db/nr) is updated with the latest protein sequences. At the same time, all the latest coordinates from the Research Collaboratory for Structural Bioinformatics (RCSB) web site (ftp://ftp.rcsb.org/pub/pdb/data/structure/all/pdb) are retrieved. Then all sequence headers in nr that have a match to a sequence from the protein coordinate database are modified to include information concerning the structures, such as resolution and number of missing atoms. The protein sequences in nr are replaced by the protein sequences actually represented in each coordinate file. In addition, we have a weekly updated fold library based on SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/). As new structures are released into the RCSB (or Protein Data Bank [PDB]), these are scanned against our existing fold library by using BLAST and subsequently added to the fold library if sufficiently nonhomologous to any existing entry.

Methods for 3D-JIGSAW

This program is modular in design with each module centering on a particular algorithm required in the modeling process. The program can either be run in fully automatic mode via a web server (http://www.bmm.icnet.uk/servers/3djigsaw), or the program modules may be executed separately and the intermediate files saved. These intermediate files can be modified, if required, before the next module in the series is run. The program modules are as follows.

Selection of parents

Parent target sequences (PDB sequences) are selected from the sequence database (described above) by using the program PSI-Blast.[6] Up to five parents are selected by using a balance of sequence similarity and data quality (e.g., resolution and number of missing atoms). The parents are then superimposed with a rigid body structure superposition program based on a pairwise superposition algorithm described by Gerstein and Levitt.[7]

Alignment of target to parent structures (approach used for CAFASP2)

Up to 50 sequences were taken from the target sequence PSI-Blast run that were evenly spaced, in terms of E values, between the target and best parent, and hierarchically aligned as described previously.[8] This multiple sequence alignment was then aligned to the multiple structural alignment of the parents. The scoring scheme here depends on the frequency of amino acids within each column of the alignments (weighted PAM250 score; modified to include gap-gap and amino acid-gap scores).

Alignment of target to parent structures (CASP4 and current 3D-JIGSAW approach)

The target sequence is submitted to a PSI-Blast run (maximum five cycles), and the position-specific scoring matrix (PSSM) is saved. The best parent sequence, selected from the PSI-Blast output, is also submitted to a PSI-Blast run (maximum 5 cycles). The target sequence PSSM is then used in the program PSI-Pred[9] to calculate the predicted secondary structure; the known secondary structure of the best template is calculated with program DSSP.[10] PSSM files from target and best parent are fed into a dynamic programming algorithm with the metric calculated within each cell of the dynamic programming matrix taken as the dot product of the target and parent PSSM at that position. This score was modified according to the local agreement of known and predicted secondary structure (addition of +1 to the log odd values if the predicted and known secondary structure are in agreement). A second dynamic program module then takes as input the alignment between the target and best template and aligns it to the multiple alignment of parent structures.

Selection of loops to change

All loops are considered for replacement. The boundaries of the loops are taken from the ends of the secondary structure elements of the multiple structure alignment. All loops and all regions with backbone angles incompatible with the target sequence were modeled via database fragment searches.[11,12]

Selection of complete backbone

From an ensemble of secondary structure elements and connecting loops a mean-field calculation is performed to select a single element or loop for all sections of the target. The algorithm used is a modification of the self-consistent mean field approach to gap closure.[13]

Selection of side-chain rotamers

Side chains are built by tracing the path of the parent side chain. After the replacement of all side chains and the assignment of a single rotamer for each, this parent rotamer plus rotamers from a side-chain rotamer library[14] are built at each residue position. A second mean-field calculation is performed to select the most probable rotamer.[15]

Energy refinement

Because the loops are modeled via database searches, they do not fit perfectly to the ends of the secondary structure elements. Thus, torsion angles were adjusted within each loop to give good geometry within the take-off regions; a modification of the tweak algorithm[16] was used for this purpose. To remove the small number of steric clashes remaining in the models, 100 steps of steepest descents energy minimization (unrestrained) were run by using the program CHARMM.[17]

Error estimates

Where possible, error estimates were made by measuring the range of equivalent atomic displacements, as measured from the superposition of all relevant homologues.[4]

Human intervention using 3D-JIGSAW

There are two critical stages in the comparative modeling process that often benefit from human intervention. The first is readjustments of the automatic alignments (see the discussion under methods for 3D-PSSM for the rationale that was also used here). The second major source of error tends to be where the end-points for loop replacements should be. Visual inspection of superimposed homologues often gives the best indication of where these are; judgments are made as to the conservation of loop stem regions, expected conformational variability within a loop, between loops, and between loops and framework.

Other much more minor interventions (e.g., adjustments to side-chain angles for side chains that consistently clash, ensuring that peptide backbone angles are within well-defined regions on the Ramachandran plot) were sometimes used; otherwise, more cycles of energy refinement would be run in the affected regions.

Methods for 3D-PSSM

Our approach to the fold recognition targets can be divided into two phases. In phase one, we initially run the target sequence through our automated server to provide us with a relatively small number of potential structural homologues. Phase two consists of a more in-depth manual analysis of these potential homologues using a variety of data sources and tools.

Phase one: Automated fold recognition

The primary method used was the program 3D-PSSM, which was described previously.[5] One of the key features of this technique is the use of multiple structural alignments of remote homologues to create extended sequence profiles (3D-PSSMs). These profiles can capture the sequence characteristics of an entire structural superfamily and extend the range of profiles generated from sequence similarity alone (e.g., PSI-Blast).

The method involves a three-pass dynamic programming algorithm against a library of known folds taken from SCOP and the PDB. Each of the three passes of dynamic programming uses sequence, secondary structure, and solvation terms. Secondary structure is matched between a known library structure and the predicted secondary structure for the query. Secondary structure prediction was performed by using PSI-Pred.[9] Our solvation model is knowledge based and similar to that reported by Jones et al.[18]

Our web server (http://www.bmm.icnet.uk/~3dpssm) reports the top 20 highest scoring structures in our library, as calculated by 3D-PSSM. In addition to the methodology described above, we have incorporated a textual component to aid in functional assignment. The program used is SAWTED[19] for Structure Assignment With Text Description (http:/www.bmm.icnet.uk/~sawted). This method compares the comments and keywords between SWISS-PROT homologues of the query sequence and the library sequence. Confident SAWTED scores are combined with the 3D-PSSM alignment scores to reflect potential functional similarity between query and template. This is intended to mimic, to a small degree, the human assessor's ability to gauge the likelihood of a correct fold or superfamily assignment based on his or her knowledge of the function of the query and template. All of the above is fully automatic and constituted our submissions to the CAFASP2 assessment, where the method proved quite successful (CAFASP2 article by Fischer et al.).

Phase two: Human intervention

Human intervention was used at two levels: (a) identification of fold and b) adjustment of the alignment.

Identification of fold. Highly confident hits from the automatic server were generally taken as correct in terms of fold. On occasion, however, such hits were rejected or modified. Clearly, in multidomain targets, confident hits would often be found to one domain and not to others. As a result, we would manually split the target accordingly and resubmit the separate domains to the automatic server. Often such determination of domain boundaries was guided by PSI-Blast multiple alignments, but it remained a rather subjective decision. In addition, when PSI-Blast had been judged to have drifted, thus bringing in spurious unrelated sequences into its profile, we would adjust PSI-Blast parameters and filtering options to limit this problem.

In cases in which no confident 3D-PSSM hits could be found for a given target, a variety of peripheral tools were used. For orphan targets, or targets with few varied homologous sequences in the sequence database, we would run tblastn at the NCBI against unfinished microbial genomes. This would sometimes supply us with both a sufficiently large and diverse set of sequences to improve secondary structure prediction accuracy, as well as a more powerful sequence profile.

When we suspected a secondary structure prediction might be erroneous, we would send the target sequence to the Jpred[20] server (http://jura.ebi.ac.uk:8888/submit.html) and look for consensus and, if necessary, run the 3D-PSSM server on a separately compiled secondary structure prediction.

When the target was present in, or homologous to, a Pfam[21] (http://www.sanger.ac.uk/Software/Pfam) family, the Pfam alignments were investigated to either generate an alternative secondary structure prediction, or to analyze alternative sequences from the same Pfam family as the target, by using the protocol above.

In addition, in difficult cases, looking at the CAFASP2 results (CAFASP2; Fischer et al.) from the many participating servers would flag up situations where 3D-PSSM had missed a target that was confidently found by other servers. In addition, the results from other servers would

focus our attention on a subset of the top 20 results from the 3D-PSSM server.

Importantly, we would make judgments about lower scoring matches from the 3D-PSSM top 20 based on the SAWTED text score and on the keywords shared between query and template. Although this feature is automatically included in the server results, below threshold SAWTED scores were often taken into account when choosing a fold or superfamily.

Alignment adjustment. Once a fold had been chosen on the basis of the above analysis, the automatic alignments produced by the 3D-PSSM server were often manually adjusted to meet a variety of criteria:

1. Maintenance of a hydrophobic core based on three-dimensional models generated from the alignments.
2. Equivalencing of known core residues (as precalculated by using a mutual contact algorithm) with hydrophobic residues in the target.
3. Preservation of the continuity of secondary structure elements.
4. Maintenance of the spatial arrangements of residues suspected to form the active site.
5. Alignment of known motifs (such as the Walker A and B motifs in P-loops, or known conserved residue types in OB folds,[22] as determined by literature searches.
6. Maintenance of the spatial distances between cysteine residues believed to form disulfide bridges.

In the relatively few cases in which we were presented with more than one high scoring, or otherwise viable template from the same fold or superfamily, we would analyze and interactively adjust the alignment to each template, looking for the template that fulfills as many of the above criteria as possible.

Finally, if very few of the above criteria could be met by any of the templates found by 3D-PSSM or other servers, an assignment of "new fold" would be submitted.

RESULTS AND DISCUSSION

Comparative Modeling

Table IA shows the levels of accuracy of the two automated programs 3D-JIGSAW (auto) and 3D-PSSM (auto) compared with 3D-JIGSAW using human intervention—3D-JIGSAW (manual). In most examples, 3D-JIGSAW (auto) performs the least well with some particularly bad RMSDs at lower levels of sequence identity between the best parent and target sequences, see, for example, T0117 and T0090. However, the original sequence alignment algorithm was not designed for alignments within the CM/FR classification (see footnote to Table I for definition of modeling sections). Column 5 of Table IA shows the results obtained by 3D-JIGSAW (auto) in October 2000, just after the CAFASP2 experiment. As described in Materials and Methods, the new alignment algorithm compares PSSMs of target and parent sequences, rather than using multiple sequence alignments. Here the results, particularly those at lower sequence identity levels, are closer to 3D-PSSM (auto) and 3D-JIGSAW (manual).

If the amount of the target modeled, and not just the lowest RMSD between model and target, is taken into account, then 3D-JIGSAW (manual) gives the better overall results. There are, however, examples of where 3D-JIGSAW (auto) (e.g., T0099 and T0125) and 3D-PSSM (e.g., T0111 and T0113) do as well. Indeed, there is only one clear example (T0128) where the manual method can be said to be far superior to the automatic methods. This particular example is described in more detail below.

For all three methods, there is no clear correlation between modeling accuracy (low RMSD) and sequence identity with the closest parent. Target T0123 (e.g., 54% sequence identity) is not modeled as well as T0122 (32%) for either the automated programs or 3D-JIGSAW (manual), average RMSD for the three methods 4.2 Å, and 2.7 Å, respectively, between equivalent Ca atoms of models and X-ray structures; actually T0123 was not modeled well by any group at CASP4. It is surprising that visual inspection of the models for T0123 show that these relatively high RMSDs are not due to alignment errors but to substantial conformational changes in loop regions between the best parent and target structure. Even more surprising is that the equivalent loops between parent and target are of the same length. It seems that the small number of sequence differences within these loops create quite different loop conformations and, thus, different packing arrangements between loops. This clearly demonstrates our lack of ability to predict, even for loops with high similarity to a parent, which loops are likely to undergo significant conformational changes between homologues.

Parent selection

Because the general comparative modeling approach can be described as essentially fitting together rigid bodies, or the fixed fragment approach, it is not possible, excluding loop regions, to model the protein backbone better than from the parent if only a single parent is used. However, this is possible if multiple parents are used in the modeling. Multiple parents were, therefore, chosen if available. However, in only one case was a model constructed that had a significantly lower RMSD than a model constructed from a single parent (T0128, 0.8 Å RMSD compared with 1.3 Å RMSD for the best single parent model).

Sequence alignment quality

Between CASP3 and CASP4 we have concentrated on algorithmic developments aimed at improving alignment quality. Although we cannot perform a statistically significant analysis on such a small sample size, we believe the following qualitative statements can be made.

All models constructed automatically above 30% identity have the correct alignment; alignments between 20 and 30% equivalence all conserved secondary structure elements but sometimes run adrift within less well-conserved elements and occasionally at the ends of the well-conserved elements; most targets below 20% sequence identity with the parent(s) showed alignment

TABLE I. Overall Accuracy of the Final Models

A. The Comparative Modeling Results

Target difficulty	Target	Sequence ID best parent (%)	RMSD 3D-JIGSAW (auto)	RMSD 3D-JIGSAW (auto) (Oct)	RMSD 3D-JIGSAW (manual)	RMSD 3D-PSSM (auto)
CM	T0128	55	1.7 (90)	1.7 (90)	0.8 (91)	1.7 (89)
	T0123	54	4.2 (99)	4.2 (99)	4.4 (100)	4.0 (98)
	T0111	51	3.0 (96)	2.3 (96)	2.3 (100)	1.8 (98)
	T0099	36	4.6 (98)	4.6 (98)	4.6 (100)	4.8 (98)
	T0122	32	2.9 (97)	2.7 (97)	2.5 (97)	2.8 (95)
	T0113	27	3.0 (93)	NC	2.4 (98)	1.9 (92)
	T0121_1	27	5.0 (62)	3.8 (100)	3.7 (100)	3.6 (98)
	T0125	18	4.7 (94)	4.7 (94)	4.7 (97)	4.8 (91)
CM/FR	T0112	24	4.1 (80)	4.1 (96)	4.0 (98)	4.2 (96)
	T0103	20	16.8 (79)	14.0 (88)	12.2 (84)	11.6 (72)
	T0117	17	13.0 (43)	NC	4.5 (79)	6.8 (70)
	T0090	16	12.0 (74)	NC	6.7 (65)	7.2 (53)
	T0089	14	21.9 (79)	19.7 (79)	17.9 (90)	17.5 (77)
	T0092	13	*	*	13.0 (85)	12.4 (78)

B. The Fold Recognition Results

Target difficulty	Target	Sippl Score 3D-PSSM (auto)	Sippl Score 3D-PSSM (manual)
FR/H	T0095_1	0	1
	T0095_2	0	4
	T0096_1	3	4
	T0098	0	2
	T0100	3	3
	T0101	2.5	2.5
	T0109	0	3
	T0110	3	3
	T0116_4	0	2.5
	T0121_2	—	2
	T0127_1	3	3
FR/A	T0102	0	—
	T0107	0	1.5
	T0108	2	2
	T0114	0	1
	T0115_1	0	0
	T0116_1	—	—
	T0116_2	2	—
	T0118	0	0
	T0120_2	3	—
	T0126	0	0
	T0127_2	1.5	—
FR/NF	T0087_2	0	0.5
	T0087_1	1.5	0.5
	T0089_2	0	0
	T0090_1	—	—
	T0091	0	—
	T0094	0	0
	T0096_2	—	—
	T0097	0	—
	T0104	0	0
	T0105	0	0
	T0106	0	—
	T0115_2	0	0
NF	T0086	0	0
	T0116_3a	0	—
	T0116_3b	0	—
	T0120_1	0	—
	T0124	0	0

†The first column shows the classification of target difficulty (defined by Manfred Sippl for CASP4). These are as follows: CM: Comparative Modeling; CM/FR: Comparative Modelling/Fold recognition; FR/H: Fold Recognition (Homologous); FR/A: Fold Recognition (Analogous); FR/NF: Fold Recognition (New Fold); and NF: New Fold. The easiest targets are classified as CM and the most difficult as NF. The targets modeled are shown in column 2. The third column of A records the percentage sequence identity between the target and best parent, calculated by using the LGA server (http://predictioncenter.llnl.gov/local/lga). Columns 4–7 record the Cα RMSDs (in Å) for all predicted residues between model and target X-ray or NMR structures for 3D-JIGSAW (auto), 3D-JIGSAW (auto) in October 2000, 3D-JIGSAW (manual), and 3D-PSSM (auto), respectively. In parentheses is the percentage of residues modeled relative to the full length of the target. The letters NC indicate that at the time this analysis was carried out, atomic coordinates for these targets were not available. The symbol * indicates where a method could not produce a model.

errors even within conserved secondary structure elements. Fortunately, most alignment errors were alleviated by human intervention above 20% identity, but a wide variety of such errors remained in the models below this level.

One major benefit of having a completely automatic comparative modeling program, modular in design, is that any one algorithm can be varied, whereas the remainder is kept fixed; the level of accuracy of the final models can then be related back to the change or enhancement of that algorithm. This we investigated by changing the alignment algorithm within 3D-JIGSAW, the before and after October 2000 algorithms, as described in Materials and Methods. Table IA indicates there is improvement particularly at the lower sequence identity level. Indeed, we believe that there is now little difference in the quality of automatic alignments between 3D-JIGSAW and 3D-PSSM. Nevertheless, although both alignment algorithms are based on the ideas of fold recognition, there are some differences in the two algorithms; for example, 3D-JIGSAW superimposes parents to create a multiple sequence parent alignment and 3D-PSSM builds up PSSMs from pairwise superposition. To test which approach is ultimately better if either (they may be sensitive to the types of protein families modeled), a larger sample size than the one shown in Table 1(A) is required. 3D-PSSM is already part of the LiveBench (http://bioinfo.pl/LiveBench/) ongoing and automatic assessment experiment. Currently, 3D-JIGSAW requires substantially more processor time than 3D-PSSM, essentially because of the three-dimensional protein fragment selection and fitting. However, once all algorithms of 3D-JIGSAW are enhanced, particularly in terms of speed, it will also be entered into the continual automatic assessments of LiveBench and EVA (http://maple.bioc.columbia.edu/eva), as also, of course, CASP5. Such experiments should help to define the finer differences in the two alignment algorithms.

Backbone modeling quality, side-chain replacement accuracy, and error estimates

Analysis of these factors shows little variation from those reported by us in CASP3.[4] Because we have not developed either the algorithms or changed the parameterization for these studies, this is not surprising.

Fold Recognition

Two questions that may be asked of our results in the fold recognition category are: What are the causes of our relative success and failure between manual and automatic techniques? Where did neither technique prove successful, that is, what are the limitations of the current method?

In Table IB, it can be seen that the vast majority of our successful predictions, both automatic and manual, arise in the FR/H (homology) section. In addition, we achieve some useful predictions in the FR/A (analogy) section. However, very little is achieved in the FR/NF and NF categories (barring some marginally useful predictions for T0087). This is to be fully expected because the 3D-PSSM technique is a homology-based technique using a library of known folds and has no ab initio or "new fold" component. As a result, henceforth, the targets in both "FR/NF" and "NF" categories will be ignored for this analysis. Although it is indeed very valuable to be able to predict the structures of such proteins, this is outside the scope of the methodology.

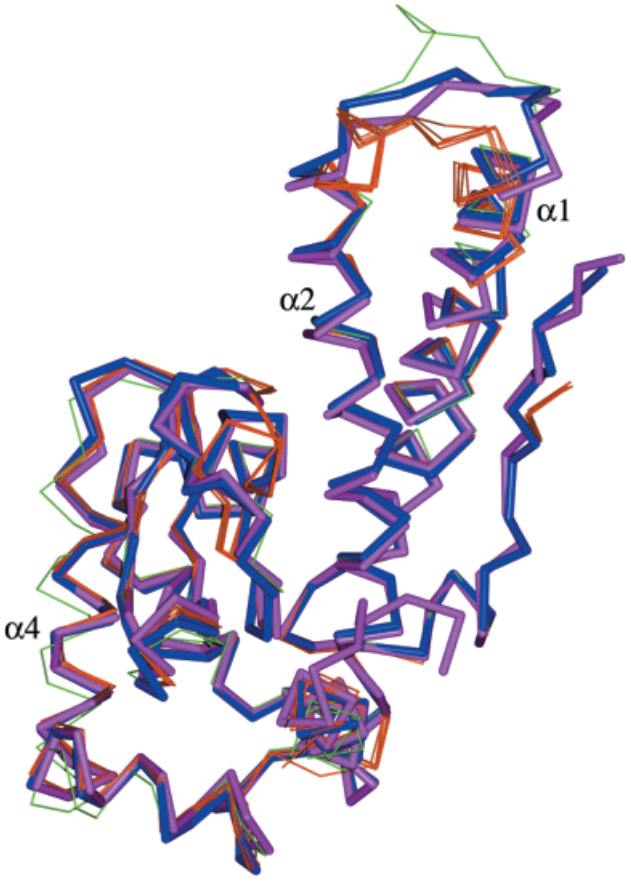

Fig. 1. Superposition of the model for T0128, manganese superoxide dismutase (*Pyrobaculum aerophilum*) (blue), generated by 3D-JIGSAW (manual), onto the first five parent structures found by PSI-Blast (orange). One of the closer parent structures to the X-ray structure (magenta) is parent 1sss_A, hyperthermophile *Sulfolobus solfataricus*) (green). This was found to be the 12th possible parent, ranked by PSI-Blast E values. The model is a composite of all possible parents superimposed in this figure. The loop between a1 and a2 in the model and X-ray structure can be clearly seen to follow a trajectory similar to the 12th parent (green). However, other parts of the model and X-ray structure follow more closely to the other parents, particularly a4.

It can be seen from Table II that, overall, the reasons for our particular successes and failures are many and varied. In general, the two major components leading to the superiority of our manual over our automatic predictions are (a) domain splitting/definition and (b) manual improvement of alignments. Our techniques for manual alignment improvement have been discussed, and it should be possible to automate these procedures to some degree in the near future. However, the correct splitting of multidomain targets into their constituent domains is an important problem to overcome. In three cases (T0116_4,T0087_2,

TABLE II. Summary of Automatic and Manual Fold Recognition Results[†]

	Reason	Number of occurences	Target ID	Score or score change from auto → manual
Where manual was superior to automatic	Manual improvement due to domain definitions	3	116_4	0 → 2.5
			121_2	0 → 2
			87_2	0 → 0.5
	Correct manual new fold assignment	2	116_3a	N/a
			116_3b	N/a
	Manual alignment improvement	2	96_1	3 → 4
			114	0 → 1
	New structure released after CAFASP deadline	1	95_1	0 → 1
			95_2	0 → 4
	Template chosen due to borderline SAWTED (text) score	1	98	0 → 2
	Using alternative algorithm/other servers	1	109	0 → 3
	Using a homologue to retrieve a confident hit	1	107	0 → 1.5
Where automatic was superior to manual	Mistaken manual new fold assignment	3	116_2	2 → 0
			120_2	3 → 0
			127_2	1.5 → 0
	Better alignment and template	1	87_1	1.5 → 0.5
General Mistakes	Skipping a domain due to a CM target	2	89_2	0
			90_1	0
	Mistaken manual new fold assignment	2	102	0
			116_1	0
	Right template, bad alignment	1	95_1	1
	Right fold, bad alignment/template	1	104	0
Equal success for both automatic and manual		4	127_1	3
			100	3
			101	2.5
			110	3
			108	2

[†]Breakdown of the reasons and resulting scores per target for the differences between the automatic and manual submissions in the fold recognition category. Sippl score refers to the assessor's scoring scheme of 0–4 (see assessor's article elsewhere in this issue).

and T0121_2) we were able to determine the presence of more than one domain within the target. For T0116_4, this was the trivial matter of noting the experimentalists' comments attached to the target. For T0087_2, we determined a better template for this domain, although we were led to this finding by the automatic method, which matched a two-domain protein for T0087 as a whole, with a similar C-terminal fold.

For T0121_2, the presence of a highly homologous domain at the N-terminus swamped our automatic results but could be detected by inspection of the alignment. In addition, there were two other cases in which a target contained domains of mixed difficulty, for example, one comparative modeling domain and one fold recognition domain; yet we were unaware of the second, more remotely homologous, domain. In the automatic technique, the relatively closely homologous domain would tend to swamp the results. In addition, because the target was not chopped into domains before submission to the server, coupled with the fact that the assessment only looks at the top ranking hit from the server, it would have been impossible for the current methodology to detect the correct fold in such cases. Because comparative modeling targets were handled separately to fold recognition targets, it was assumed no such remotely homologous domain would be present in the target.

In contrast to our manual improvements, the automatic technique sometimes fared better. On three occasions where we judged a target to adopt a new fold because of a lack of confident scores, the top ranking automatic 3D-PSSM hit was found by the assessors to be correct to varying degrees. We cannot yet accurately differentiate new folds (or FR/NF targets) from analogues and homologues with borderline or insignificant scores. This is evident from the fact that we erroneously labeled three targets as "new fold," and conversely, submitted several models for targets that were in the FR/NF or NF categories. It should be noted that without knowing before the commencement of the CASP experiment whether a correct submission of "new fold" will be rewarded or whether an erroneous fold assignment will be penalized, it is not possible to make a rational decision as to one's policy of submission.

It was also apparent that the more time was allowed for a prediction, the greater the likelihood of either sequences or structures being released to the public, which could aid in an accurate fold assignment. For target T0095, this is most clearly demonstrated, because a relatively close homologous structure was released in the PDB after the automatic CAFASP deadline but before the manual CASP deadline. In several other cases, it was observed that the confidence of certain 3D-PSSM matches increased over the course of the CASP experiment, once again because of database updates.

Fortunately, there were no cases of false positives (i.e., confident E value assigned to an incorrect fold or superfam-

ily), although for target T0126, we achieved multiple hits to the same fold (all subthreshold), and this is usually considered as evidence of a correct hit, although this turned out to be erroneous.

Finally, it should be pointed out that small all-alpha proteins and long helical targets (T0124) are particularly difficult to recognize because of the lack of signal from patterns of secondary structure and problems of PSI-Blast drift, respectively.

CONCLUSIONS

It is most important for the CASP experiments to summarize what went right and what went wrong with our current approach. The following steps seem to be movement in the correct direction.

1. The use of multiple parents, if selected and superimposed carefully, can produce models with higher accuracy than from any single parent (e.g., T0128).
2. Both automatic programs seem to be selecting the correct parents and producing alignments of reasonable quality above 20% sequence identity with the current algorithms.
3. The regular and automatic updating of sequence and structural databases contributes considerably to detection of both remote homologues and for comparative modeling enables a diverse selection of high-quality parents.

There are however several major shortcomings.

1. Automatic alignments could be improved below 20%.
2. Database searches cannot find good candidates for long loops or model links between domains.
3. Need for an automatic technique for splitting target sequences into domain.

Over the years, comparative protein modelers have been criticized for only doing well at high levels of sequence identity with a single parent where they simply copy much of the backbone and side-chain conformers. Although there remains an element of truth in this statement, there are clear signs that models can now be built that are more accurate than any built from a single parent (e.g., T0128). However, such models cannot be built routinely. Two major problems remain here. When we do copy from parents, which bits are the more reliable? If we have created a model better than any parent, how do we know that we have achieved this? Finally, how can we achieve this enviable situation for all comparative models, even if only one possible parent exists?

The distinction between targets amenable to comparative modeling and fold recognition is beginning to blur. As comparative modeling moves into the "twilight zone," it will become progressively more useful for modelers to adopt the techniques of fold recognition to aid in both parent selection and accurate alignment. It is tempting to speculate that as these trends continue, the entire range of protein structure prediction, from comparative modeling, to fold recognition and "new fold" predictions, will be tackled by using a common set of tools: multiple sequence profiles and secondary structure matching for both remote homology detection and for high-accuracy alignments for comparative modeling; fragment packing techniques for both novel fold building, as well as loop modeling in close homologues.

REFERENCES

1. Sali A, Sanchez R. Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. Proc Natl Acad Sci USA 1998; 95:13597–13602.
2. Jones TA, Kleywegt GJ. CASP3 comparative modeling evaluation. Proteins 1999;Suppl 3:30–46.
3. Murzin AG. Structure classification-based assessment of CASP3 predictions for the fold recognition targets. Proteins 1999;Suppl 3:88–103.
4. Bates PA, Sternberg MJE. Model building by comparison: At CASP3: Using expert knowledge and computer automation. Proteins 1999;Suppl 3:47–54.
5. Kelley LA, MacCallum RM, Sternberg MJE. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:501–522.
6. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
7. Gerstein M, Levitt M. Using interactive dynamic programming to obtain accurate pairwise and multiple alignments of protein structure. Fourth International Conference on Intelligent Systems for Molecular Biology. Menlo Park, CA: AAAI Press; 1996. p 59–67.
8. Barton GJ. Sternberg MJE. A strategy for the rapid multiple alignment of protein sequences: Confidence levels from tertiary structure comparisons. J Mol Biol 1987;198:327–337.
9. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
10. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983:22:2577–2637.
11. Jones TA, Thirup S. Using known substructures in protein model building and crystallography. EMBO J 1986;5:819–823.
12. Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJE. An automated classification of the structure of protein loops. J Mol Biol 1997;266:814–830.
13. Koehl P, Delarue M. A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. Nat Struct Biol 1995;2:163–170.
14. Tuffery P, Etchebest C, Hazout S, Lavery R. A new approach to the rapid determination of protein side-chain conformations. J Biomol Struct Dynam 1991;8:1267–1289.
15. Koel P, Delarue M. Application of a self-consistent mean field theory to predict side-chain conformation and estimate their conformation entropy. J Mol Biol 1994;23:249–275.
16. Shenkin PS, Yarmush DL, Fine RM, Wang H, Levinthal C. Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. Biopolymers 1987;26:2053–2085.
17. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy minimization and dynamics calculations. J Comp Chem 1983;4:187–217.
18. Jones DT, Taylor WR, Thornton JM. A new approach to fold recognition. Nature 1992;358:86–89.
19. MacCallum RM, Kelley LA, Sternberg MJE. Bioinformatics 2000; 16:125–129.
20. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. Jpred: A consensus secondary structure prediction server. Bioinformatics 1998;14:892–893.
21. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EEL. Pfam 3.1: 1313 multiple alignments match the majority of proteins. Nucleic Acids Res 1999;27:260–262.
22. Bycroft M, Hubbard TJP, Proctor M, Freund SMV, Murzin AG. The solution structure of the S1 RNA binding domain: A member of an ancient nucleic acid-binding fold. Cell 1997;88:235–242.