PROTEINS: Structure, Function, and Genetics, Suppl. 1:29–37 (1997)

Homology Modeling With Internal Coordinate Mechanics: Deformation Zone Mapping and Improvements of Models via Conformational Search

Ruben Abagyan,* Serge Batalov, Timothy Cardozo, Maxim Totrov, Jeremy Webber, and Yingyao Zhou
Skirball Institute of Biomolecular Medicine, Biochemistry Department, New York University Medical Center, New York, New York

ABSTRACT **Five models by homology containing insertions and deletions and ranging from 33% to 48% sequence identity to the known homologue, and one high sequence identity (85%) model were built for the CASP2 meeting. For all five low identity targets: (i) our starting models were improved by the Internal Coordinate Mechanics (ICM) energy optimization, (ii) the refined models were consistently better than those built with the automatic SWISS-MODEL program, and (iii) the refined models differed by less than 2% from the best model submitted, as judged by the residue contact area difference (CAD) measure [Abagyan, R.A., Totrov, M.J. Mol. Biol. 268:678–685, 1997]. The CAD measure is proposed for ranking models built by homology instead of global root-mean-square deviation, which is frequently dominated by insignificant yet large contributions from incorrectly predicted fragments or side chains. We demonstrate that the precise identification of regions of local backbone deviation is an independent and crucial step in the homology modeling procedure after alignment, since aligned fragments can strongly deviate from the template at various distances from the alignment gap or even in the ungapped parts of the alignment. We show that a local alignment score can be used as an indicator of such local deviation. While four short loops of the meeting targets were predicted by database search, the best loop 1 from target T0028, for which the correct database fragment was not found, was predicted by Internal Coordinate Mechanics global energy optimization at 1.2 Å accuracy. A classification scheme for errors in homology modeling is proposed. Proteins, Suppl. 1:29–37, 1997.** © 1998 Wiley-Liss, Inc.

Key words: deformation zones; prediction map building; homology modeling

INTRODUCTION

High-quality modeling by homology remains extremely important, given the order of magnitude difference between the number of proteins with experimentally determined three-dimensional structure and their relatives for which only the sequence is available. Modeling by homology consists of alignment to the best 3D template (or templates), mapping of the possible zones of backbone deviation, and, finally, placement of side chains and the mapped tentative loops by a database search[1–4] or a global energy optimization.[5–10] An error incurred at any of the steps of homology modeling is unrecoverable at the subsequent steps. Therefore the early steps of model building have a disproportionally large influence on the net result of modeling and should be more carefully studied and developed.

The ICM method[11–13] attempts to globally optimize the energy of all-atom, arbitrarily constrained models with respect to free internal variables. The energy includes the vacuum energy of an all-atom model as well as solvation, surface and side-chain energy terms.[9,14] ICM global energy optimization of the initial model includes side-chain prediction effected with the Biased Probability Monte Carlo method,[14] in which continuous random steps are performed according to the multivariate angular probability distributions, and loop prediction in a flexible protein environment.[9,12] The latter method correctly predicted two seven-residue loops *before* the structures were determined crystallographically.[8,15]

In this article, we analyze the results of our blind predictions for the five low-sequence identity and one high-sequence identity homology modeling targets (T0001, T0003, T0009, T0017, T0024, T0028) for which the answers were made available. First, we show that a new procedure, prediction map building, should be introduced and developed because knowing the exact loop boundaries before conformational search is critically important: loop limits may be

Abbreviations: CAD, contact-area difference; ICM, Internal Coordinate Mechanics; NCS, noncrystallographic symmetry; RMSD, root-mean square deviation; 3D, three-dimensional.

Contract grant sponsor: Department of Energy; Contract grant number: DE-FG02-96ER62268.

*Correspondence to: Dr. Ruben Abagyan, Skirball Institute of Biomolecular Medicine, Biochemistry Department, NYU Medical Center, 540 First Avenue, New York, NY 10016.

Received 16 May 1997; Accepted 25 August 1997

asymmetrical with respect to the insertions/deletions, and large local deviations can occur even in ungapped parts of the alignment. Second, we answer three questions: (i) were initial models improved by the ICM energy optimization including side-chain placement and loop prediction? (ii) were the refined models better than the models built by the automatic SWISS-MODEL server,[16] and (iii) was there at least one loop predicted correctly by global energy optimization but not with any other method (there was none in CASP1)? Fortunately, the answer to all three questions is positive. We analyze what went right and wrong with the models; a new measure for evaluating models is proposed, and we suggest a method for predicting local backbone deformations.

BUILDING MODELS BY HOMOLOGY

To build the models we (A) chose one or several templates; (B) aligned sequences to their templates; (C) mapped the tentative backbone deformation zones, and, finally, (D) predicted side chains, loops, and termini by the ICM global energy optimization. In this article we introduce an additional step (C) of prediction map building in which the sequence-structure alignment is divided into alternating zones of structurally conserved regions and "deformation zones." Structural information for the former is directly inherited from the template, while the structure of the latter is predicted in step D.

We used the Needleman and Wunsch alignment with a blosum45 residue exchange matrix,[17] zero gap-end penalties and normalized gap opening/extension penalties of 2.8 and 0.15, respectively[18,19] to align trial sequences to their template sequences. From the total of 23 insertions and deletions (further referred to as indels) in our sequence alignments for all five targets, four, in T0001, T0003 and T0028, were manually edited to new positions. Two deletion regions and one insertion region were moved from a relatively extended template fragment to a point of higher curvature, typically the tip of a beta-hairpin, and one deletion zone was moved to preserve a disulfide bridge. In retrospect, all four alignment decisions resulted in correct residue assignments.

The third (C) modeling step, prediction mapping, was not automated. There are three typical situations requiring a prediction map decision.

1. An isolated indel. Question: where are the loop boundaries? One needs to decide precisely where the deformation zone around the indel position in the alignment will begin and end.
2. Two or more indels are close to each other, and/or sequence similarity in the alignment fragment between them is weak. Question: should one merge the two tentative loops into one big loop?
3. Weak alignment in a nongapped region of an alignment or at the chain end. Question: should one create a new deformation zone?

These decisions were made on the basis of the three-dimensional structure of the template, structural variability regions observed in a superposition of all available homology templates (targets T0001 and T0003), and/or positions of strong sequence identity around a tentative deformation zone. For each of 23 cases of indels we had to make a prediction map decision. In about two thirds of the cases, we used the first flanking residue identity as a signal of deformation zone termination. An example of an exception to this rule: in the first loop of T0009 the conserved hydrophobic substitution V → F was used as the left boundary, because the first identity was too far away (this decision turned out to be correct). In target T0028 three indels between 280 and 299 were correctly merged into one deformation zone because of the weak sequence similarity in the region. We chose no ungapped deformation zones.

The fourth modeling step, side-chain prediction, was automated and performed as previously described.[9,14]

The fifth modeling step, the prediction of backbone deformations (loop prediction) was also automated. In retrospect, only two out of the 23 indels were noninteracting loops with correctly predicted boundaries! Therefore only these two had a chance of accurate prediction by energy optimization, although local features of several loops were correctly determined despite the distorted environment. Of these two, loop1 in T0028 was predicted with 1.2 Å RMSD upon superposition of the five residue loop stems (Table I). The final modeling step, energy minimization of the backbone, was only attempted for target T0017. The minimization moved the model away from the template and the right answer.

ACCURACY OF THE SUBMITTED MODELS

Traditionally, the accuracy of models built by homology has been evaluated by global cartesian RMSD from experimental coordinates.[20,21] However, this measure is dominated by fluctuations in the incorrect and/or insignificant parts of the structure such as wrongly predicted loops/ends of rearrangements of long exposed side chains. Furthermore, global changes dominate over cumulative local similarities, for example, two ideally predicted, but shifted, domains may have RMSD from the correct structure which is worse than the two correctly positioned domains with totally incorrect internal structure. CAD, a new robust geometrical measure to evaluate models by homology, proposed recently,[22] reflects the degree of restoration of interresidue contact areas. It is sensitive to both side chain and backbone conformations and is applicable in a wide range of model accuracy. This measure was calculated via the areas of interresidue contacts, A_{ij}^{R} and

TABLE I. Most Accurate Loop Predictions on Homologous Templates and the Two Insertion Loops of T0028*

Target	Loop res.	*N*	Indel type	Flanks (N + C)	*L*	Sternberg RMSD	Moult RMSD	Cohen RMSD	Abagyan RMSD	Best method
T0028	76–81	0	Del	2 + 2	4	**0.48**	0.51	0.56	2.55	Database search
T0028	239–244	0	Del	2 + 2	4	**0.85**	2.01	1.78	2.35	Database search
T0028	332–338	0	Del	2 + 2	4	**0.86**	0.94	2.90	2.43	Database search
T0028	214–221	0	Del	2 + 2	4	**1.13**	1.27	1.32	4.23	Database search
T0028	42–48	0	Del	2 + 3	5	1.46	3.13	2.01	**1.19**	Energy optimization
T0028	96–104	3	Insert.	3 + 4	10	9.46	7.71	7.18	8.09	All methods were incorrect
T0028	256–268	8	Insert.	4 + 1	13	12.53	5.74	11.12	11.40	All methods were incorrect

*Listed in the top section above are the five best predicted loops sorted by the RMSD accuracy of the best prediction. The next best prediction has RMSD of 1.58 Å. RMSD was measured for the loop Cαs, upon superposition of five flanking Cα atoms on each side of the loop. Loop res. lists the consensus loop definition produced by the overlap of all the predictors' loop definitions. *N* is the number of inserted residues (0 for deletions). In/Del indicates whether the loop was marked by inserted (In) or deleted (Del) residues. Flanks (N + C) indicates the number of residues on the respective ends of the inserted residues which, with knowledge of the solution structure, would have been optimal flanking residue selections. *L* indicates the length of the resulting ideally defined loop (= N + Flanks). Numerical results are in angstroms, and best predictions are bold and underlined. All of these loops turn out to be from target T0028, which was also the target with the highest sequence identity template among those which contained gaps (loops). These loops are also all deletion loops. The corresponding values for the only two insertion loops in this target are given at the bottom of the table for comparison.

A_{ij}^{M}, in reference structure R and model M, respectively:

$$\mathrm{CAD} = C\,\frac{\sum_{i,j} |\, A_{ij}^{R} - A_{ij}^{M}|}{\sum_{i,j}(A_{ij}^{R} + A_{ij}^{M})}$$

where i and j were residue numbers and scaling factor $C = 180\%$ was used to evaluate modeling by homology. The contact areas A_{ij} were calculated with a probe sphere of 1.4 Å and the following van der Waals radii: C, 1.9 Å; N, 1.7 Å; O, 1.4 Å; S, 1.8 Å. Hydrogen atoms were ignored. As one can see if conformations R and M are identical, all the differences $|A_{ij}^{\mathrm{R}} - A_{ij}^{\mathrm{M}}|$ are equal to zero and CAD = 0%. If two conformations are totally unrelated CAD is close to 100%.

The CAD measure gives a more objective, and totally automatic, comparison between CASP models without the exclusion of clearly wrong fragments. If a fragment loses its correct contacts, its contribution to CAD simply becomes zero regardless of where the wrong contacts are formed. Situation is the opposite for RMSD, which becomes dominated by highly fluctuating contributions from the incorrect fragments. For example, models of the T0028 target submitted by the Cohen and Moult groups are virtually identical by the CAD measure with the Cohen group model being marginally better (24.81% and 25.03%, respectively). However, the Moult model has a significantly lower RMSD from the solution (4.02 Å and 3.37 Å for Cohen and Moult groups, respectively) (Fig. 1). In reality, this difference results from three *entirely wrong* large loops for *both* models (e.g., the best prediction for one of these loops, residues 256:268, is 5.74 Å from the solution) (Table I). If these three loops (176:190, 256:268, 280:299) are removed, the RMSD values move to 2.09 Å and 2.14 Å, closely reflecting both the order and the degree of difference of the CAD comparison before removal.

Figure 1 shows both the CAD measures and all-heavy-atom RMS deviations between the correct answers and the submitted models. The participating groups are specified in the legend of Figure 1. We submitted models for all six targets and in each case there was only one final model.

In all five low-sequence identity models our final model was within 2% CAD from the best model submitted, the CAD range for other models being 25%, 25%, 40%, 15%, and 11% for T0001, T0003, T0009, T0024, and T0028, respectively. All ICM models were within the best 10% of the above ranges, as measured by CAD. In T0009 the model was best by both RMSD and CAD, the difference with the second best model being 7% in CAD and 0.65 Å in RMSD. This target was the lowest sequence identity comparative modeling target. In T0001 and T0024, ICM models were among three best models with nearly indistinguishable CAD measures (0.7 and 0.3%, respectively). For target T0028, the ICM model was inferior to the Cohen and Moult models by 2% CAD and for target T0003 was 2% CAD worse than the Sali model. The initial model for high sequence identity target T0017 was in the group of seven models (including the SWISS-MODEL) of indistinguishable quality (within a range of 1% in CAD and 0.1 Å in RMSD). However, when this model was freely energy-minimized it moved away from the template and the crystallographic answer by about 2% in CAD and 0.3 Å in RMSD, which made it worse than its own initial model and the SWISS-MODEL (by 0.5% CAD).

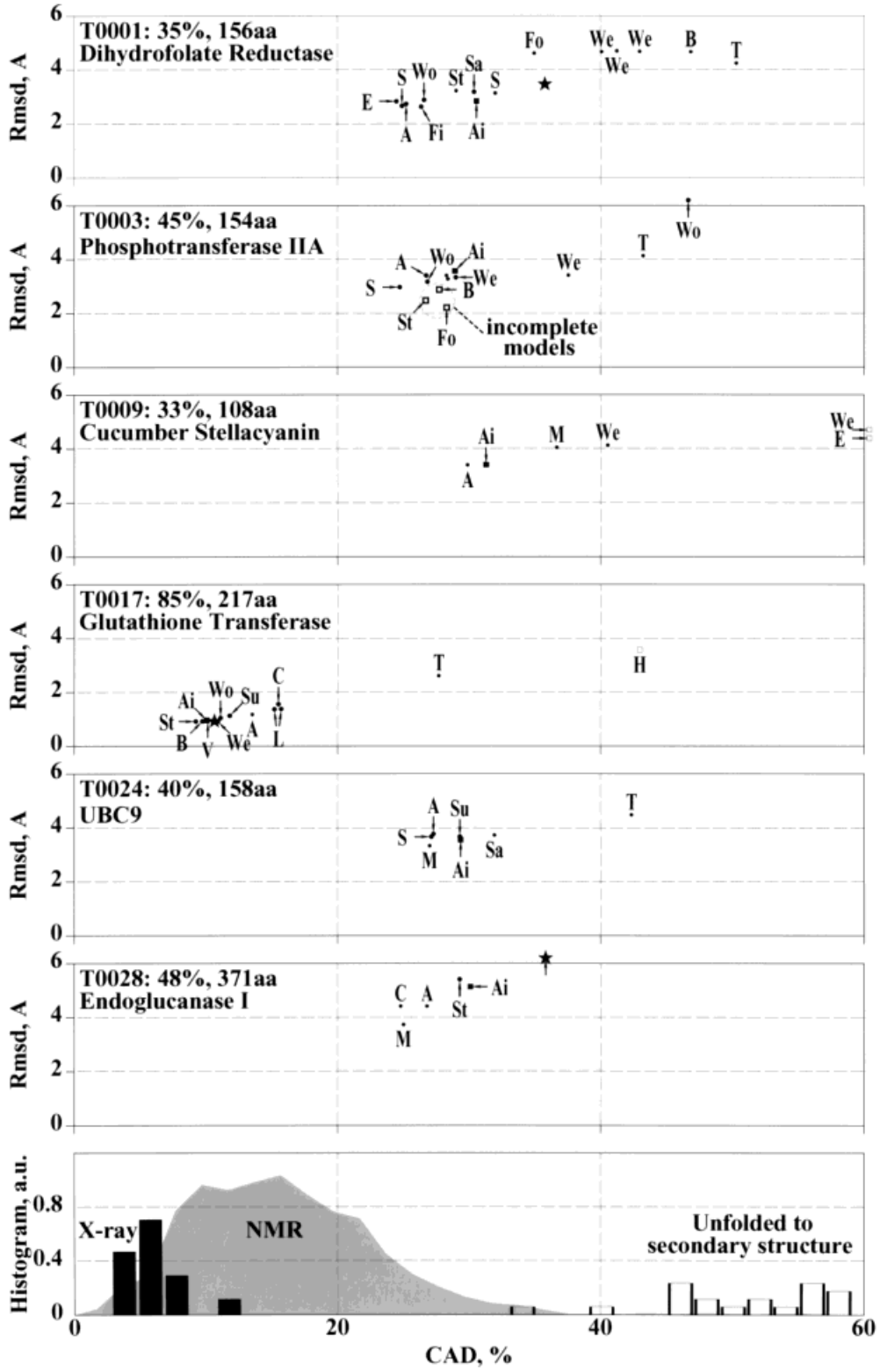

Figure 1 (legend on following page).

For *all* five low identity models the ICM global energy optimization of both side chains and loops *improved* the *initial* models. T0001 was improved by 5% CAD, T0003 by 2% CAD, T0009 by 2% CAD, T0024 was improved by 2% and T0028 by 3%.

Four of five final low-identity targets were also substantially better than SWISS-MODELs by 11%, 36%, 6%, and 9% CAD for T0001, T0009, T0024, and T0028 targets, respectively. T0003 had to be truncated to the SWISS-MODEL size to be compared fairly and was slightly worse by 1% CAD. Furthermore, SWISS-MODEL automatically truncated from 8 to 38 of the most difficult residues, whereas we submitted complete models in all cases.

We collected all loops in the submitted models and compared Cα RMSD after superimposition of the five-residue loop stems. Only a few relatively short deletion loops were predicted correctly (Table I). ICM global energy optimization predicted the best loop 1 in T0028 while the best loops for four other deletion loops were predicted by the Sternberg group using a database search method[23] (the Moult and Cohen groups had similar accuracy for these four loops). No insertion loops were predicted correctly. What was the accuracy of the best models by homology compared to the accuracy of experimental methods of structural determination, such as X-ray crystallography and NMR? In contrast to the cartesian RMSD, the CAD measure reveals a substantial deterioration of structural quality in models by homology in the 33–48% identity range. The best models in this range are close to the worst 10% of NMR models in the PDB and are far beyond the range of natural structural plasticity caused by different crystal environments, which is exhibited by the pairs of NCS-related dimers. Even the 85% identical no-gap T0017 model only approaches the natural plasticity limit of about 5–9% CAD, the best model having 9.24% CAD with the experimental structure.

Fig. 1. Overall distribution of the prediction results for the six targets on the CAD measure and the RMS deviation of heavy atoms (equivalent atoms were iteratively found in symmetrical side chains). The groups of predictors are denoted as follows: A, Abagyan; Ai, Abagyan (initial models, solid squares); B, Bruccoleri; C, Cohen; E, Egner; Fi, Fidelis; Fo, Forster; H, Honig; L, Lee; M, Moult; S, Sali; Sa, Saqi; St, Sternberg; Su, Sutcliffe; T, Taylor; V, Vriend; We, Weber; Wo, Wolynes. The *open squares* mark the incomplete predictions (6 residues omitted by Sternberg and Bruccoleri groups and 9 by Forster in T0003, 14 by Egner, Weber in T0009). The *asterisks* mark automated predictions from SWISS-MODEL server[16] if they were complete (T0001, T0017, T0028). For T0003 and T0024 the SWISS-MODELs were lacking 16 and 38 residues, respectively, and the SWISS-MODEL for target T0009 was built on the basis of low quality Ca template, while the CASP2 predictors used a full atom template. The histogram at the bottom shows the distribution of the CAD differences between all high resolution identical NCS-related domains determined by X-ray crystallography, differences between alternative models in NMR submissions to the Brookhaven database, and the CAD changes upon unfolding into secondary structure elements.[22] The low accuracy of the T0017 model from the Honig group was due to scientific honesty, the predictors believed that using 85% identical template would have been cheating, and they used a low sequence identity template.

WHAT WENT WRONG AND WHY

The problems in modeling by homology can be classified into the following categories (we use the same nomenclature as in our classification of the homology modeling steps): (A) inadequate choice of the template or templates; (B) misalignment; (C1) unexpected region of structural deviation in an ungapped part of the sequence-structure alignment; (C2) incorrect boundaries of the deformation zone around a gap in the alignment; (D1) intolerable distortion of structural environment of side chains or loops; (D2) distortion of loop stems is too large; (D3) loop conformation is influenced by another undefined loop or chain end which is ignored by the search procedure which assumes the environment rigid; (D4) loop is too large for a prediction by both database method and energy optimization; (D5) the energy function is too inaccurate. In our predictions, the most problematic were error types were C1 and C2 which have never been properly addressed previously. A list of problems in the six submitted models classified into the above categories follows.

We had one case of *misalignment* (error B) of a fragment of eight residues in T0003 near the C terminus. We trusted local sequence similarity of four identical residue pairs in a continuous stretch of eight residues (the two stretches were QEDIVKI and QGDVVAI, the template and target, respectively), which turned out to be misleading. The lesson from that misalignment is that the chain ends can deviate from the template position despite a very high sequence similarity.

We chose no ungapped deformation zones (error C1). This decision was incorrect in at least two cases. Most of the local fragments forming a subdomain in target T0001 between residues 60 and 100 deviate with regard to the best available template. The packing in this subdomain is altered enough to warrant marking at least some of the fragments as ungapped deformation zones. In target T0009, the beta strand at (66,72) deviated strongly from the template. This ungapped area of the alignment DNDVERT/GAKVYTS was not defined as a deformation zone, since the preceding zone was already too large.

Could one possibly predict such a deviation? The recently derived statistics for structural significance of sequence alignment[19] could give a helpful hint just on the basis of local sequence information (Fig. 2). (C). In target T0003, we defined two loops at residues 131 and 146 near the C terminus and did not merge them (incorrectly). In Target T0009, two indels between 48 and 61 were merged to form one deformation zone with the helical region in between remaining locally as a mobile rigid body. This decision was incorrect: the low similarity region (61:74)

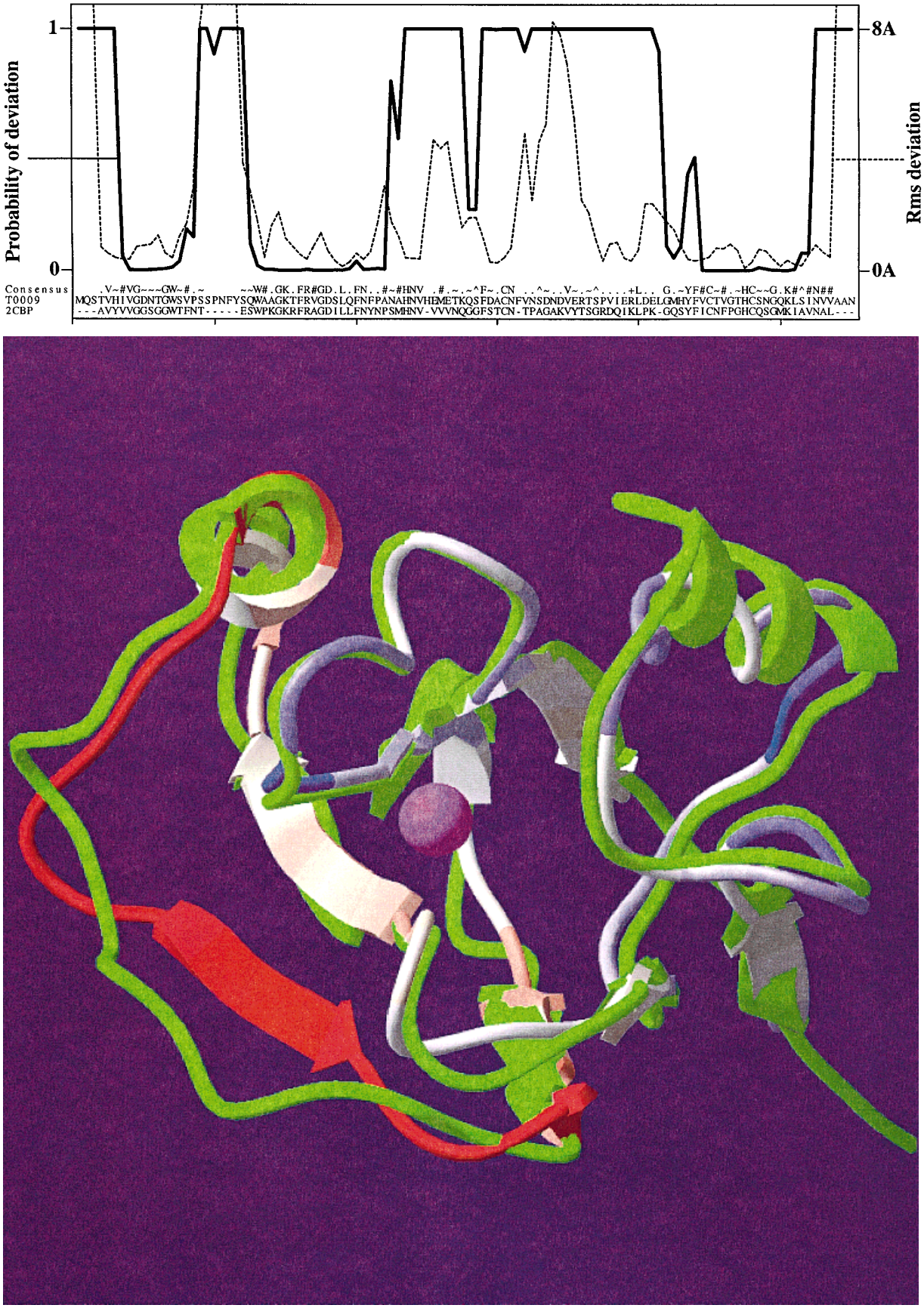

Figure 2 (legend on following page).

C-terminal to this fragment deviated between the homologous template and the experimental structure.

In T0024, the loop-1 definition was too large; loop 2, the boundary defined by C-terminal flanking sequence W × P, was too short. Actually, the alignment of tryptophans should have been ignored. For the two insertion loops of T0028, the loop boundaries defined by the nearest sequence identity rule were insufficient. The actual deformation was five residues longer than expected in both cases.

We encountered numerous occurrences of error types D1–D5 in the predictions of side chains and loops. Finally, energy minimization of the whole model, including the backbone, which was mistakenly applied to the high sequence similarity model T0017 resulted in some deterioration of the model. We conclude that the whole backbone should not be optimized without tight restraints.

RECOMMENDATIONS FOR BEGINNERS

The recipe is simple:

1. Template choice. Take the closest sequence similarity template,[9] keep the backbone *exactly* as it is in the template throughout all the steps and *never* energy-minimize the inherited backbone. Combine (but do not average) the template with fragments from other homologous structures if they are closer in sequence locally.
2. Alignment. Start with a sequence alignment. Review the alignment very carefully, since in most cases it is the main determinant of the model quality. The main rules are the following:
   MEM rule (make ends meet): Shift deleted template fragments (deletion loops) to make their ends as close as possible (e.g. make them reside symmetrically at the tip of a β-hairpin).
   NCI rule (no core insertions): Shift insertion location from the compact core to the surface.
   EAT rule (ends are tricky): N and C terminal fragments may be misaligned despite sequence similarity.
3. Prediction map. Carefully decide which backbone regions can deviate from the template(s). The rules: Fan rule: Divergent regions (fan-like) in a visual superposition of multiple templates may prompt zone limits.
   GUFI rule (go until first identity): Apply to determine deformation boundaries about both insertion and deletion loops.
   LLSS rule (low local sequence similarity): If an LLSS alignment fragment is near an indel or between two indels and is on the surface, merge the fragment with the indel(s) GUFI zones into one deformation zone.
   EAT rule (see above): you may need to define the entire terminal fragment (usually until the first indel) as a deformation zone despite sequence similarity, provided the end is not buried in the core.
4. Place all nonidentical side-chains in the most statistically probable conformation, and check rotamers for clashes, e.g. ref. 24, although the latter is not going to improve the previous placement much, at least in terms of RMSD.
5. Stop here and drop all the deformation zones from the model. The only hope for improving the model is to search for four-residue deletion fragments in the database.

The model is ready for submission. Any further manipulation will reduce its accuracy.

Fig. 2. **A (top):** Correlation of the probability of local structural difference calculated in a window of 11 residues using the analytical formulae for structural significance of sequence alignment[19] (*bold line*) with the RMS deviation of the backbone Cα atoms between the template and the experimentally determined structure for T0009. Molecules were superimposed according to the optimal structural alignment. The large deviation of the beta strand (66,72) could have been predicted by this method. The probability of structural deviation was calculated as follows: (1) all 11-residue stretches of the sequence alignment shown were extracted; (2) for each alignment fragment the alignment score S was calculated with the Blosum45[17] residue substitution matrix and gap penalties of 17.5 and 0.94; (3) probability P of structural deviation was calculated as

$$P = 1 - e^{-e^{-1.618y}}$$

where

$$y = \frac{t - (2.24 + 0.006L)}{1.15 + 0.0014L}$$

and L is the window size of 11 residues. The consensus line uses # sign for conserved hydrophobic residues, ∧ sign for conserved small residues and ∼ sign for conserved polar residues. **B (bottom):** The same reliability function shown by color (*blue,* reliable; *red,* unreliable) on a structural superposition of our model of T0009 and the experimental structure (*green* ribbon).

DISCUSSION

Modeling by homology is a complicated multistep process in which many important decisions are made even before the conformational search is begun. Although we found that our conformational searches improved overall geometrical similarity of models to their template(s) by 2–5% of CAD in all low sequence similarity cases, these presearch steps, involving choice of template, sequence-structure alignment and, particularly, deformation zone mapping, turned out to be the main contributors to the accuracy of low- and medium-sequence similarity models. CAD was essential in revealing the consistent pattern of improvement caused by the ICM global optimization, since the RMSD analysis of the side chains and loops is obscured by dominating errors incurred in the pre-search steps.

Previously we concluded that choosing the closest template[9] locally might be preferable to template averaging.[3] Only one target (T0001) in CASP2 needed

a chimeric template in which the gapped alignment with the main template was replaced by an ungapped alignment fragment with a homologous protein. Combining an optimal chimeric template from many templates deserves an automated algorithm.

Alignment, the next step in the pipeline, remained critically important, but it is the subsequent step, the deformation zone determination, which caused most of the problems. Three alignment rules used in this work, the MEM rule for deletion loops, the NCI rule for insertion loops and the EAT rule for chain ends (see previous section), were sufficient to get sequence-structure alignment right in all but one case, C terminus of T0003, in which the EAT rule was underestimated.

Prediction of the exact limits of loops and other zones of backbone deformation in most cases goes beyond simple addition of two flanking residues to the gap in the alignment or the first flanking identity (GUFI) rule. Here we show how a dramatic deviation of beta strand 66:72 in T0009 could have been anticipated from an analysis of statistics of local sequence alignment score (Fig. 2). An ideal future method would combine the sequence significance signal with the structural characteristics of the template.

An important result of our analysis is that we found a consistent pattern of improvements of initial models via global energy optimization of side chains and loops. In all five low sequence similarity targets the CAD to the correct answer was reduced from 2% to 5%. However, this improvement becomes visible only if the CAD measure[22] is used instead of RMSD, which is dominated by variations of contributions from obviously incorrect model parts.[22]

The overall improvement of models is consistent but small, partially due to the fact that no single insertion loop was correctly predicted (similarly to CASP1). Although one deletion loop was predicted reasonably well by energy optimization (Table I), there are simply too many problems with loop determination (the determination of correct loop boundaries; deformed structural environments; shifts of loop stems; interactions between loops etc.: see errors C2,D1–D5 described above). Although for many loops convergence of the ICM global optimization was achieved and the accuracy of energy function seemed sufficient, the mentioned problems made correct prediction impossible. One way to account for deformations of the structural environment of a loop is to include the surrounding side chains and potentially surrounding loops into calculation. Another way of "softening" the environment in a simulation could be the use of smoothened grid potential instead of the explicit atomic model to represent the structural environment of the loop.

Finally, unrestrained energy minimization of the backbone can only make a model worse because the correct conformation is clearly unreachable by a simple local minimization or a short run of molecular dynamics, and, furthermore, potential functions are too inaccurate for such a delicate operation even if one starts from the correct crystallographic structure.

The new CAD measure was used here for structure comparison. In comparison with cartesian coordinate RMSD, the CAD measure is somewhat more difficult to compute and it weakly depends on the choice of atomic van der Waals radii and the water probe radius. However, the CAD measure solves the principal problem in comparison of partially correct models, since it measures the fraction of the correct features of the model, rather than focusing on how wrong are the wrong parts of the models, as it was shown above for to models of t0028. The CAD measure is automatically insensitive to the movements of the surface side chains, adequately sensitive to domain rearrangements and protein plasticity, and fairly ranks the models in a wide range of model accuracy. For example, in this paper we showed that even the best models between 33% and 48% of sequence identity are far beyond the average accuracy of NMR structures. Only the 85% sequence identity models of t0017 reached NMR accuracy.

CONCLUSIONS

Modeling by homology is a sequential process of (i) alignment of a query sequence with the template; (ii) mapping regions which can deviate substantially from the template; (iii) prediction of side chains, and (iv) prediction of loops. An error at each step is unrecoverable by later procedures, therefore early steps have larger practical importance.

Prediction map building is important because it precedes prediction of side chains and loops. Automated methods to predict the extent of local deviations around gaps and in the weak un-gapped parts of the alignment should be developed. Structural significance statistics based on the alignment score[19] can be used to predict unexpected local deviations. Chain ends can be shifted despite high local sequence similarity.

Side chains predicted with the Biased Probability Monte Carlo method consistently improve the initial models.

Distorted environments, neighboring loops or incorrect boundaries were the most frequent causes of incorrect loop prediction. Loop 1 (43:47) in T0028 with its relatively undisturbed environment was best predicted by the ICM global energy optimization procedure (1.2 Å accuracy). Four other successful "deletion loops" were best predicted with a straight database search. No "insertion" loops were predicted correctly. Prediction of loops in a "soft" environment by allowing rearrangements of the neighboring side chains and, potentially, even backbone movements of surrounding backbone fragments may improve performance.

With regard to the energy minimization of the whole model we conclude, that the exact backbone of the template should be used in structurally conserved regions. Energy minimization of the model backbone may cause the model to move away from the correct answer.

CAD measure is better than cartesian RMS deviation as an integral measure of the prediction quality for models by homology.

ACKNOWLEDGMENTS

We thank Department of Energy (grant DE-FG02-96ER62268) for financial support. DOE's support does not constitute an endorsement by DOE of the views expressed in the article.

REFERENCES

1. Jones, T.A., Thirup, S. Using known substructures in protein model building and crystallography. EMBO J. 5:819–822, 1986.
2. Chothia, C., Lesk, A.M. Canonical structures for the hypervariable regions of immunoglobulins. J. Mol. Biol. 196:901–917, 1987.
3. Blundell, T., Carney, D., Gardner, S., et al. Knowledge-based protein modelling and design. Eur. J. Biochem. 172:513–520, 1988.
4. Levitt, M. Accurate modeling of protein conformation by automatic segment matching. J. Mol. Biol. 226:507–533, 1992.
5. Moult, J., James, M.N.G. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. Proteins 1:146–163, 1986.
6. Bruccoleri, R.E., Karplus, M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 26:137–168, 1987.
7. Bruccoleri, R.E., Haber, E., Novotny, J. Structure of antibody hypervariable loops reproduced by a conformational search algorithm. Nature 335:564–568, 1988.
8. Borchert, T.V., Abagyan, R.A., Kishan, K.V.R., Zeelen, J.Ph., Wierenga, R.K. The crystal structure of an engineered monomeric triosephosphate isomerase, monoTIM: The correct modeling of an eight-residue loop. Structure 1:205–213, 1993.
9. Cardozo, T., Totrov, M., Abagyan, R. Homology modeling by the ICM method. Proteins 23:403–414, 1995.
10. Vasquez, M., Nemethy, G., Scheraga, H.A. Conformational energy calculations on polypeptide and proteins. Chem. Rev. 94:2183–2239, 1994.
11. Mazur, A.K., Abagyan, R.A. New methodology for computer-aided modelling of biomolecular structure and dynamics. 1. Non-cyclic structures. J. Biomol. Struct. Dyn. 6:815–832, 1989.
12. Abagyan, R.A., Mazur, A.K. New methodology for computer-aided modelling of biomolecular structure and dynamics. 2. Local deformations and cycles. J. Biomol. Struct. Dyn. 6:833–845, 1989.
13. Abagyan, R.A., Totrov, M.M., Kuznetsov, D.A. ICM: A new method for structure modeling and design—Applications to docking and structure prediction from the distorted native conformation. J. Comp. Chem. 15:488–506, 1994.
14. Abagyan, R.A., Totrov, M.M. Biased Probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J. Mol. Biol. 235:983–1002, 1994.
15. Thanki, N., Zeelen, J.Ph., Mathieu, M., Jaenicke, R., Abagyan, R.A., Wierenga, R.K., Schliebs, W. Protein engineering with monomeric triosephosphate isomerase (monoTIM): the modelling and structure verification of a seven residue loop. Protein Eng. 10:159–167, 1997.
16. Peitsch, M.C. ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. Biochem. Soc. Trans. 24:274–279, 1996.
17. Henikoff, S., Henikoff, J.G. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U.S.A. 89:10915–10919, 1992.
18. Needleman, S.B., Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48:443–453, 1970.
19. Abagyan, R.A., Batalov, S.V. Do aligned sequences share the same fold? J. Mol. Biol. 273:355–368, 1997.
20. Moult, J., Judson, R., Fidelis, K. A, Pedersen, J. T. A large-scale experiment to assess protein structure prediction methods. Proteins 23:ii–iv, 1995.
21. Mosimann, S., Meleshko, R., James, M.N.G. A critical assessment of comparative molecular modeling of tertiary structures of proteins. Proteins 23:301–317, 1995.
22. Abagyan, R.A., Totrov, M. Contact Area Difference (CAD): A robust measure to evaluate accuracy of protein models. J. Mol. Biol. 268:678–685, 1997.
23. Bates, P.A., Sternberg, M.J.E. In Rees, A.R., Sternberg, M.J.E., Wetzel, R. "Protein Engineering: A Practical Approach." Oxford: IRL Press, 1992:117–141.
24. Vriend, G. WHAT IF: A molecular modeling and drug design program. J. Mol. Graphics 8:52–56, 1990.