Homology Modeling by the ICM Method

Timothy Cardozo, Maxim Totrov, and Ruben Abagyan
Skirball Institute of Biomolecular Medicine, Biochemistry Department, NYU Medical Center, New York, New York 10016

ABSTRACT **Five models have been built by the ICM method for the Comparative Modeling section of the Meeting on the Critical Assessment of Techniques for Protein Structure Prediction. The targets have homologous proteins with known three-dimensional structure with sequence identity ranging from 25 to 77%. After alignment of the target sequence with the related three-dimensional structure, the modeling procedure consists of two subproblems: side-chain prediction and loop prediction. The ICM method approaches these problems with the following steps: (1) a starting model is created based on the homologous structure with the conserved portion fixed and the nonconserved portion having standard covalent geometry and free torsion angles; (2) the Biased Probability Monte Carlo (BPMC) procedure is applied to search the subspaces of either all the nonconservative side-chain torsion angles or torsion angles in a loop backbone and surrounding side chains. A special algorithm was designed to generate low-energy loop deformations. The BPMC procedure globally optimizes the energy function consisting of ECEPP/3 and solvation energy terms. Comparison of the predictions with the NMR or crystallographic solutions reveals a high proportion of correctly predicted side chains. The loops were not correctly predicted because imprinted distortions of the backbone increased the energy of the near-native conformation and thus made the solution unrecognizable. Interestingly, the energy terms were found to be reliable and the sampling of conformational space sufficient. The implications of this finding for the strategies of future comparative modeling are discussed.**

Key words: modeling by homology, protein structure prediction, loop modeling, side-chain placement, Monte Carlo procedure

INTRODUCTION

The meeting on the Critical Assessment of Techniques for Protein Structure Prediction (CATPSP) represents the first attempt to standardize evaluation of various modeling methods objectively. This exercise provides data on real in which the prediction of different side chains and new loops is severely complicated by unknown and complex distortions in backbone and conserved side chains of the template. To date only a handful of such "real" cases of homology modeling have been published as test cases in which the prediction was then evaluated against the solution, most for side-chain prediction[1–11] and only two for divergent loops of the protein.[12,13,43]

The proteins modeled in this work (Table I) represent a wide range of sequence identities, and each therefore emphasizes a different step of the modeling procedure: alignment, threading, side-chain placement, loop prediction, and global refinement. Traditional Needleman and Wunsch[14] pairwise alignment provides a reasonable starting alignment. The next step may take the three-dimensional structure into account. There are many automated methods of threading (reviewed in 15,16) which are tested by their ability to recognize the fold from a set of candidate structures. The quality of each individual alignment in these automations may be rather poor.[17,18] Therefore, for the single threading of sequence onto a homologous structure, expert adjustment by eye supported by multiple sequence alignment would give the best alignment in most cases. Targets of moderate to high sequence identity (>35%) generally provide enough information, especially with regard to the core, to continue the modeling procedure.[19]

Another aspect of building the starting model is the decision of which homologous structure, if there are several, will be used in the model. Depending on the structural diversity, an average of several chains may be useful.[20] The alternative is choosing the single one closest in sequence identity or a hybrid structure of the most closely related fragments patched together.

With the starting structure chosen, the placement of side chains becomes the issue. This task has been presented as a combinatorial problem of staggering magnitude.[6,8] To some extent, the problem is ameliorated by the reliable assumption that conserved

Received April 13, 1995; revision accepted June 5, 1995.

Address reprint requests to Ruben Abagyan, Skirball Institute of Biomolecular Medicine, Biochemistry Department, NYU Medical Center, 540 1st Avenue, New York, NY 10016.

TABLE I. CATPSP Meeting Targets Modeled with ICM*

Prediction target	Protein (size in residues)	PDB homology	Sequence identity (%)
p450eryf	Cytochrome (404)	1CPT	24.8
crabp	Cellular retinoic acid binding protein (136)	2HMB	41.8
hpr	Histidine-containing phosphocarrier (89)	2HPR	45.9
E5.2 antibody	Iodotype antibody	6FAB	62.8 (heavy), 88.8 (light)
nm23h2	Nucleoside diphosphate kinase (151)	1NDL	77.5

*The three main tasks of homology modeling short of global refinement are represented.

side chains retain the same conformation as in the homologue.[21] More importantly, recent studies have shown that the problem is overstated and that the backbone to a large extent determines an individual side-chain conformation with little or no coupling between side chains.[1] As such, side-chain placement has been accomplished, with varying degrees of success in the past few years.[1–11,21,22] The assessment in the CATPSP meeting of this step of the modeling task may therefore be particularly illuminating in defining the state-of-the-art.

Finally, with an understanding that global refinement of a model including the backbone does not bring structure closer to the solution with the current technology, the obstacle of comparative modeling at present is loop modeling. Both database and conformational search methods may be applied to the determination of the loop. Database search methods have been used to generate starting conformations for the loop[23] as well as to model the loop.[24] Conformational search procedures used include systematic search (usually CONGEN),[12,23,25,26,45,46] molecular dynamics,[27,28] Monte Carlo simulated annealing,[29,30] and Internal Coordinate Mechanics (ICM).[31] The coordinates of the loops ends, determined by the threading step, may be particularly important as boundary conditions. Indeed, a major issue is whether the modeling of loops can procedurally be uncoupled from the computationally expensive global refinement steps. If the answer is yes, it may be possible to model the loop. If no, loop modeling may become as complicated a task as ab initio prediction of the whole protein and would require a modification of the current approach.[32]

ICM was used to model the proteins for the CATPSP meeting. This is an efficient method for global optimization of ECEPP/3 energy combined with a solvation energy term and modeling restraints with respect to an arbitrarily chosen set of internal variables such as torsion angles. The rest of the model is fixed. The program's capability to exclude large numbers of variables, rather than restrain them, and make rationally designed moves allows drastic increases in the sampling efficiency.[33] The method has been successfully used for ab initio peptide structure prediction,[33] side-chain prediction,[1] protein design and loop prediction,[13] and biomolecular docking.[34] For the meeting targets, we focused on the side-chain placement and loop modeling components of the comparative modeling procedure, presenting two of the targets (E5.2 antibody.predict and crabp.predict) in depth as an example of the loop modeling problem and two others (nm23h2.predict and hpr.predict) as examples of the side-chain prediction problem.

This paper presents the results of the modeling and analyzes the results. The following issues are addressed: Does it make sense to model low sequence identity targets beyond the threading step? What criteria decide the best choice of homologous structure? Is side-chain placement truly too large a combinatorial problem to be accurately accomplished? Is the sampling of the procedure extensive enough for loop modeling? Is the energy function sufficiently accurate? Can the loop modeling step be isolated from global refinement including the backbone?

METHODS

Identification of homologous proteins in the Protein Data Bank was the first step in the method, followed by threading of the target sequence onto the homologous structure or structures, building of the initial model, placement of side chains, and loop prediction. Global refinement was not attempted.

Threading

The initial pairwise sequence alignment[14] was done for the target and its nearest homologue. For targets nm23h2 and hpr, this step was sufficient (determined by visual inspection) and no further threading was necessary. In the other three cases, this initial alignment was adjusted manually to ensure that the insertions and deletions occurred in the turns rather than in the middle of secondary structure elements. Graphical adjustment was satisfactory in the case of E5.2 antibody and most of the loops in p450eryf, however with the two loops in crabp and the rest of the loops in p450eryf, we had to decide to which side of a β-strand or a helix the insertion or deletion was to be moved. To provide additional clues for this decision, all the homologous sequences were extracted from the SWISSPROT database and all the homologous structures extracted from the Protein Data Bank (PDB). A multiple sequence and structure alignment was then performed by the ICM 2.0 program. The multiple sequence alignment algorithm is similar to the that of CLUSTAL and its extensions.[35–37]

Building the Starting Model

The starting model for both side-chain placement and loop prediction contains all atoms (including hydrogen atoms), exhibits idealized covalent geometry throughout, and takes the conformation of the aligned parts of the backbone and conserved side chains from the nearest related solved structure. The model is built with an extension of the sequential regularization procedure described in Abagyan et al.[31] After the regularization step, the model was relaxed by a series of iterations in which the vacuum energy and the distance restraints (tethers) to the aligned atoms of the nearest homologous structure were minimized with respect to all the torsion angles in order to relieve the steric strain. The weight of the tether penalty term was reduced from iteration to iteration as the conformation was relaxed. At the end of the procedure, polar hydrogens were placed by a systematic search.[31]

Side-Chain Placement

To predict the conformations of all nonconservative chain chains, the Biased Probability Monte Carlo (BPMC) procedure was applied to the starting model in the following strategy:

1. Side-chain torsion angles were freed for all nonconservative side chains, all the other variables were fixed.
2. Preferred statistical zones were loaded and assigned to the nonconservative side chains.[33]
3. The Biased Probability Monte Carlo minimization procedure was applied to predict all the side-chain conformations simultaneously by global optimization.[33]
4. The best energy conformation was subject to energy minimization allowing small movements in the backbone from the homologous coordinates.

The objective function optimized at steps (3) and (4) contained ECEPP/3[38–40] and solvent-accessibility-based solvation energy.[41] The combined energy was optimized according to the double energy scheme.[31] Typically the energy plots showed convergence after 200,000 energy evaluations. One run took from 16 to 36 h on the Indigo 2 workstation.

Loop Prediction

The ends of the loop in antibody E52 were assigned to the terminal residue of the preceding and to the original residue of the following B-strand as determined by a combination of visual inspection and pairwise sequence alignment. The third heavy-chain loop (third hypervariable region, H3) was clearly defined by multiple sequence alignment as the only loop containing an insertion or deletion (1 residue insertion) and therefore the only one which would definitely require conformational search. With the idea of restricting the conformational space as much as possible without excluding a possible solution, visual consideration of the threading and the high sequence identity within the loops led us to leave the L1-3 and H1-2 loops of E5.2 antibody fixed. In the case of crabp, the gap for each loop was placed by pairwise sequence alignment in the center of an α-helix for loop I or β-strand for loop 2. Multiple sequence and structure alignment was then used to identify the locally divergent region and the loop was moved to that area (the C-terminal side in both cases). In all three loops, the modeling zone was defined as the residues of the loop itself and all the substituted side chains within 6 Å of any loop residue. Tethers to the homologous structure were applied to all the remaining atoms of the model. A loop deformation procedure[42] was applied to generate conformations.

RESULTS

Threading of Actinomyces p450eryf (p450eryf)

This target was attempted only as a check of our threading method which would provide the first step for the subsequent model and is dealt with briefly here. The closest homologous structure was 25% identical in sequence and the solution was by X-ray crystallography. Our prediction and the starting structure had C_α coordinates 2.2 Å rmsd. The solution C_α was 4.4 Å from the prediction and 4.6 Å from the starting structure. The true test of correct threading is determining which of the elements of secondary structure present in the solution are represented in the homologous structure and how many of these we placed correctly in our prediction. Comparison of the locations of these elements in the prediction and the solution shows that most residues were correctly assigned (Fig. 1). However, the rmsd data illustrate the futility of pursuing optimization of the threaded model at this sequence identity. No current procedure can hope to converge to the correct minimum from an rmsd greater than 2 Å for a protein of this size, and indeed, for much smaller proteins. Previous work has demonstrated convergence for less than 2 Å in an ideal case indicating that there is at least hope for this range of rmsd distortion.[31]

Side Chain Prediction of Human Nucleoside Diphosphate Kinase (nm23h2)

This protein was solved by NMR as a hexamer and therefore provides six solutions for each substituted side chain. The known homologous structure had 77% identity to the sequence. Figure 2 shows the predicted conformation with the six solutions superimposed on the backbone. At issue is how the solution data were generated and, thus, how clustering of the solutions should be interpreted. In addition,

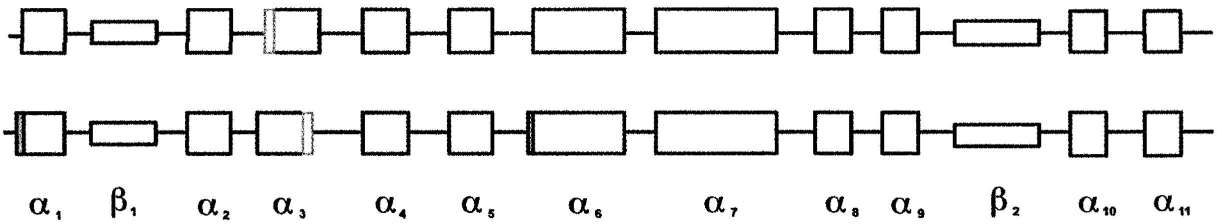

Fig. 1. Residue assignment within secondary structure elements conserved from the starting homologue (1CPT/2CPP hybrid) to the native conformation. There is an extra residue assigned to helices 1 and 3 and helix 3 is translocated by 1 residue.

structures solved by NMR show an inverse relationship between the range of possible conformations and the solvent accessibility of the residue. With these factors in mind, we define a correct prediction as one which falls within the spread of provided solutions. This can be assessed visually for each individual side chain (Fig. 3). Most fall within the spread of solutions and in those that do not appear to (i.e., Arg-34, Lys-66, Lys-124), the spatial distribution of the six NMR solutions is wide, suggesting that the side chain is flexible in solution and that there is no single correct conformation. Correspondingly, Figure 2 displays nonpolar and polar side chains separately. It is clear that the polar side chain solution conformations vary much more widely than the nonpolar ones, except in the case of the aromatic residues. This is an indication that the existence of a single conformation for many free, polar side chains on the surface of the protein may be a fallacy.

Side Chain Prediction of Mycoplasma Histidine Containing Phosphocarrier Protein (hpr)

The closest homologous structure was 46% identical in sequence and the solution was by X-ray crystallography. Table II lists the results of the prediction for the each of the substituted side chains. We used our function to calculate the energy of the native side-chain conformation in the setting of our model. Presumably, if our energy function is reliable, the native conformation, if it is different, will lower the energy of the protein and our conclusion would be that we did not cover the conformational space adequately. If our sampling procedure was effective, however, the energies should be comparable and visual inspection of the side chain should reveal an identical conformation. If the energy of the native conformation is higher, the conformation is different, and no explanation can be found for the discrepancy, the energy function is called into doubt. The results of the energy calculations were that none of the near-native conformations, when placed in the environment of the model, results in a lower energy. Thus, our sampling was sufficient. Figure 4 displays the individual side chains. Of the 47 32 are correct by visual inspection and rmsd (not shown) alone. Of the remaining 15, only 4 (Ile-7, Leu-14, Thr-57, Thr-84) have *B*-factors (Table II) indicating sharp electron density. The others have high *B*-factors which may account for the conformational discrepancy. (Indeed, all contain charged groups (Glu, Asp, etc.) and are on the surface of the protein indicating that they may not have a single conformation.) Of these four residues, two, Ile-7 and Leu-14, exhibit local backbone movements which altered the template for the prediction. Ile-7 interacts with the four C-terminal residues of the protein which contains a one residue addition (Gly-89) and is displaced from the starting structure coordinates. Leu-14 exhibits a local backbone shift at residues 14–16. Since side-chain conformation depends mostly on the backbone,[1] these shifts mean that the template for the prediction of these two side chains was not accurate causing the energy of the near-native conformation to be higher and thus unrecognizable. The remaining two residues with higher near-native energies and different conformations are both polar threonines located on the surface of the protein. The water molecules appearing in the solution are too few to represent those actually present and thus it is likely that this difference is due to the geometry of the packed water moleules in the crystal interacting with the oxygen of threonine. ICM models the protein in a virtual solution environment and therefore finds a different minimum.

Thus, we find no side chain with a sharp coordinate solution exhibiting a different conformation and a lower energy. We are also able to identify the cause of the higher energy near-native state in each of the remaining side chains leading us to conclude that both our energy function is reliable, and that our sampling was sufficient to correctly place the side chains.

Loop Modeling of Mouse Cellular Retinoic Acid Protein (crabp)

There were two proteins in this prediction target, mouse and human species. We attempted only the mouse structure. The closest homology was 42% and the protein structure was determined by X-ray crystallography. In both loops of the CRABP target, transposition of the near-native coordinates to the model results in a higher energy (Table III). As such, the ICM modeling algorithm would not have selected the solution conformation. The reason for the

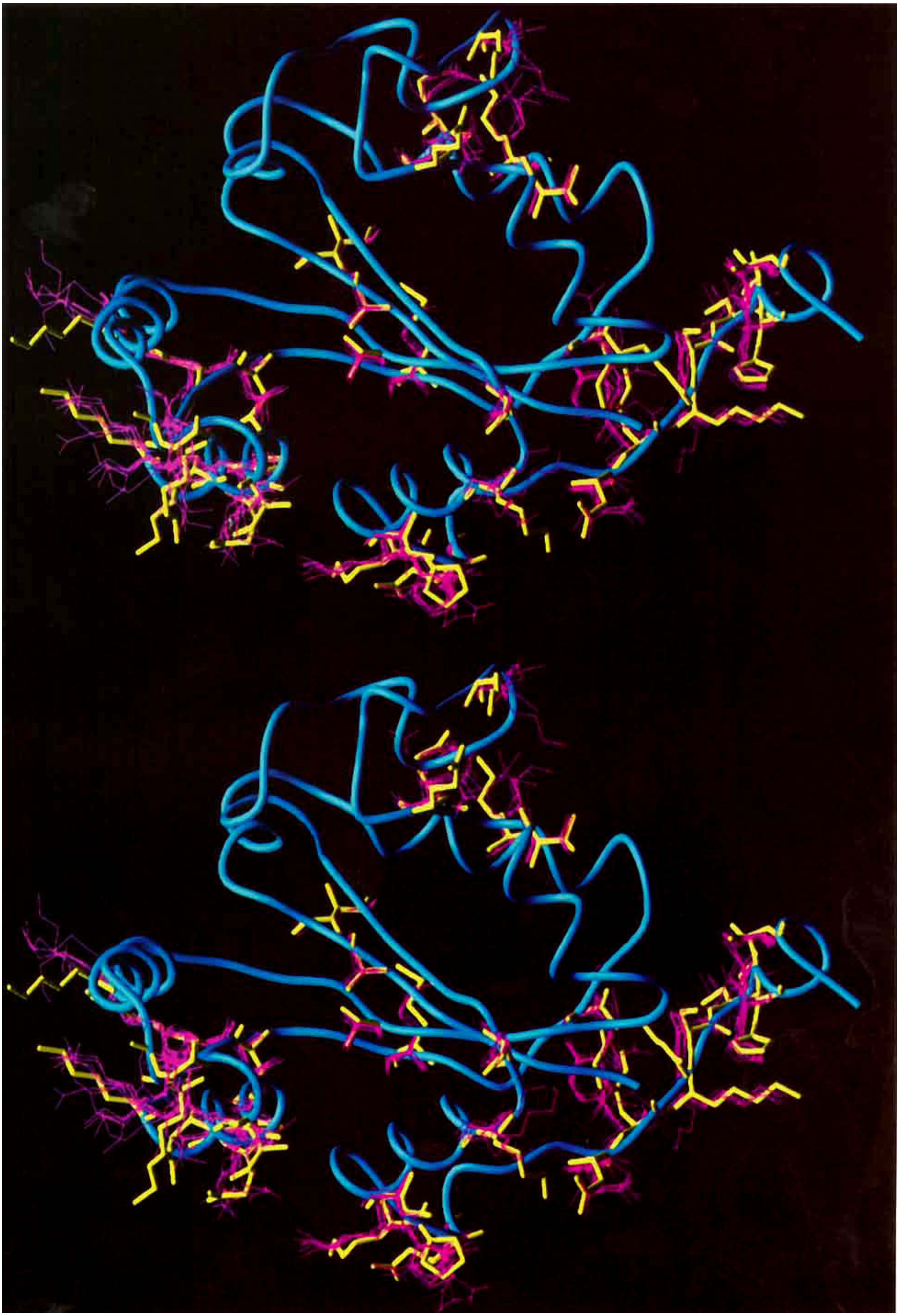

Fig. 2. Superimposition of the six NMR solutions of nm23h2 side chains (magenta) on the prediction (blue ribbon, common backbone; yellow stick, predicted side chains). Only the predicted nonconserved side chains are shown. The conserved side chains always correspond perfectly and are omitted for clarity.

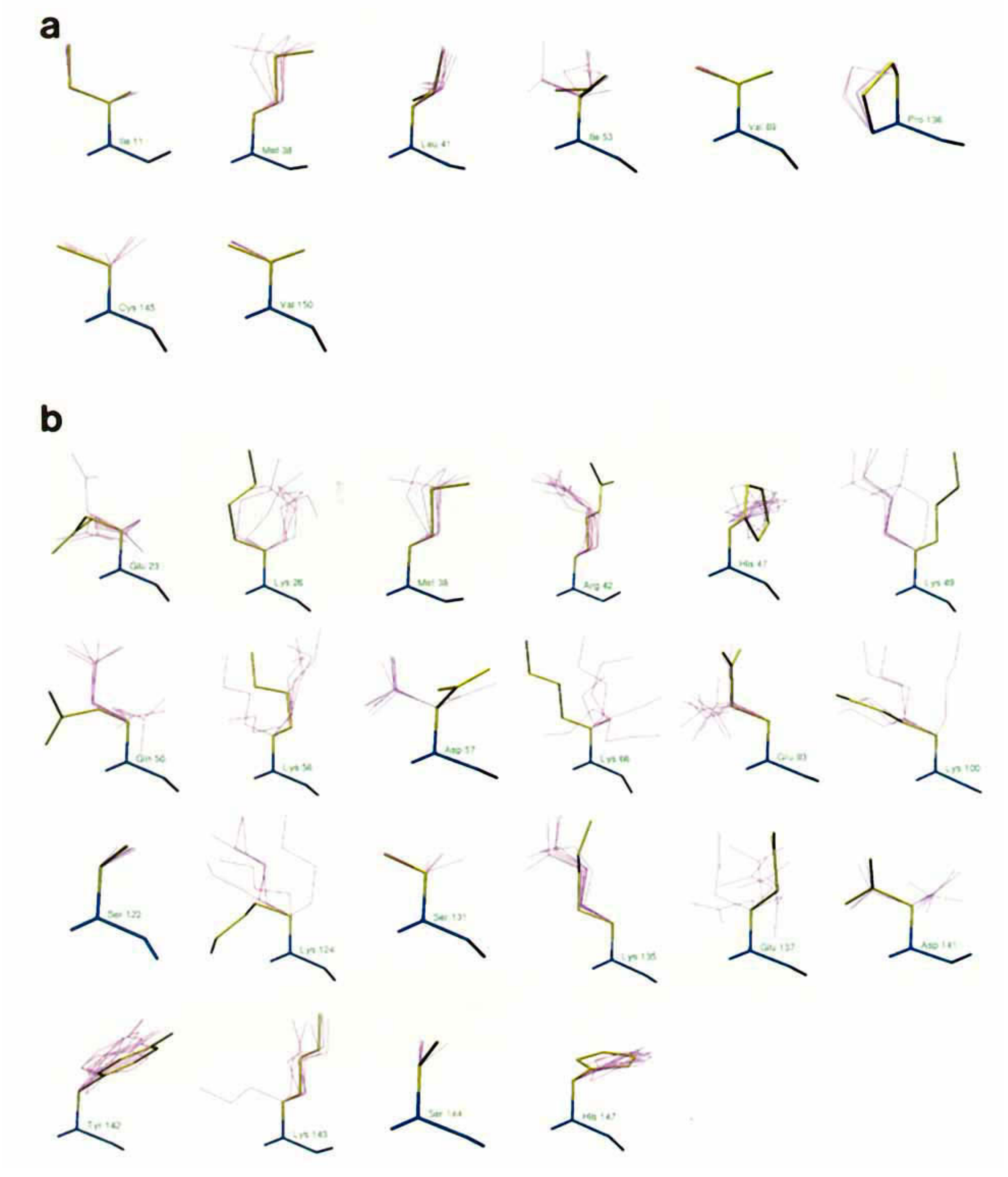

Fig. 3. (**a**) Nonconserved side chains of the nm23h2 NMR solution (magenta) superimposed on the ICM prediction (yellow stick). Nonpolar residues are shown. The backbone (blue) is oriented so that all views are comparable with the C-terminus to the right. (**b**) Side chains of the nm23h2 NMR solution (magenta) superimposed on the ICM prediction (yellow stick). Polar residues are shown. The backbone is oriented with the C-terminus to the right. (Note: Lys-135 is a sequence error, the solution is residue is Glu.)

discrepancies become clear upon careful inspection of the surrounding environment of the loops: Backbone differences between the template and the solution are large and clearly invalidate the modeling parameters (Fig. 5). The loop at residues 34–39 is located at the C-terminal of a distinctive helix-turn-helix feature of the protein. The rest of the protein is essentially a sandwich of two β-pleated sheets so

TABLE II. Results of Side-Chain Prediction for All the Nonconserved Side Chains in hpr*

Residue	Result	Residue	Result	Residue	Result
Lys-3	Correct	Asn-32	High *B*-factor	Gln-64	High *B*-factor
Ser-5	Correct	Ile-33	Correct	Asp-66	High *B*-factor
Ile-7	**Backbone/crystal**	Thr-34	High *B*-factor	Asn-68	High *B*-factor
Ile-8	Correct	Ile-35	Correct	Asp-71	High *B*-factor
Asp-10	Correct*	Ile-36	Correct	Gln-72	High *B*-factor
Lys-11	Correct*	Glu-39	High *B*-factor	Ile-74	Correct*
Val-12	Correct	Gln-41	High *B*-factor	Gln-75	High *B*-factor
Leu-14	**Backbone/crystal**	Asn-49	High *B*-factor	Ile-77	Correct*
Ser-20	Correct	Met-51	Correct	Lys-78	Correct
Lys-24	High *B*-factor	Met-53	Correct	Gln-79	Correct
Glu-25	Correct	Lys-56	Correct	Ile-82	Correct
Ser-30	Correct	Thr-57	**Backbone/crystal**	Asp-83	Correct
Ser-31	Correct	Lys-59	Correct	Thr-84	**Backbone/crystal**

*Only four with low *B*-factors display a different conformation in the solution as opposed to the prediction (bold). The backbone or water molecules in the crystal influenced the prediction (see text). The others are either correct or have high *B*-factors. A few (*) have high *B*-factors only for the atom(s) which diverge from the prediction (see Fig 4b: Asp-10, Ile-74, Ile-77).

that the helix-turn-helix emerges from a β-strand and ends at a β-strand with few or no intervening residues. Insertion of two residues to form the loop results in extension of the C-terminus helix of a highly conserved helix–loop–helix fold by one-half turn. In order to accommodate this turn, the entire helix-turn-helix is displaced obliquely and the direction of its long axis relative to the rest of the protein changed in order to pack correctly (Fig. 5a). Our parameters did not allow this movement (correctable perhaps by "loosening" of a hinge variable at the beginning of the second helix).

Similarly, the loop at residues 100–108 containing a four residue insertion adopts a conformation dominated by extension of a β-strand to Leu-101 and alignment of Leu-101/Glu-102 with the plane of a β-sheet. This conformation is afforded by hydrogen bonding with Lys-81 which is positioned appropriately due to a backbone movement in the β turn at residues 77–81. Accordingly, the energy calculations show a greater contribution of hydrogen bonding terms to the distortion than in the previous loop, although the van der Waals strain is still dominant (Table III). Residues 77–81 are exactly conserved in sequence from the homologous structure—their repositioning could not have been predicted without considering all the other conserved residues. In modeling both loops, our procedure found a minimum from 4 different starting conformations. Building a representative model from the solution and standardizing the energy terms to those used in the modeling run allow evaluation of the native conformation energy as it would be calculated in the ideal case of a perfectly conserved template (Table III, last column). The data reveal that the native conformation represents lower van der Waals and hydrogen bonding energy for both loops. Thus, imprinted differences in the backbone template contributed significantly to energy evaluation of loops and obscured the solution.

In both these cases, the conserved backbone movements were large and might be addressed by breaking up the template (which was considered a single rigid body in the modeling) into movable blocks. This might allow a limited and practical form of simultaneous global refinement. However, as the next model illustrates, it is not at all certain that this limited global refinement will be sufficient to allow accurate loop prediction.

Loop Modeling of E5.2 Antiiodotype Antibody to D1.3 (antilysozyme) Antibody (E5.2 Antibody)

The homologous structure had a 89% sequence identity in the light chains and a 64% sequence identity in the heavy chains. The solution was by X-ray crystallography. The near-native energy of this loop was higher as well, but by less than in the previous case (Table III). The difference in the van der Waals strain was about 350 kcal, but the absolute value of this energy for the near-native structure was closer to zero. This is an indication of the closeness of the boundary conditions of the loop conformational space to the native structure (loop ends, surrounding backbone). Indeed, superimposition of the model on the solution reveals no visually obvious shifts in the backbone template (data not shown). In fact, one of our assumptions—that use of the nearest identity structure rather than an average of structures provides a better template for the prediction—is borne out by the closeness of the light chain chosen to the solution light chain (Fig. 6). Calculating the energy of the solution with the perfect template once again shows an improvement in van der Waals (Table III). However, in this exemplary case of a real loop modeling situation, we have shown that even provided the best template possible (avoiding the large backbone movement of the previous cases), established reliability of the energy function and adequacy of sampling,[13] the template is still not sufficient to allow accurate prediction of the loop due to small van der Waals imperfections

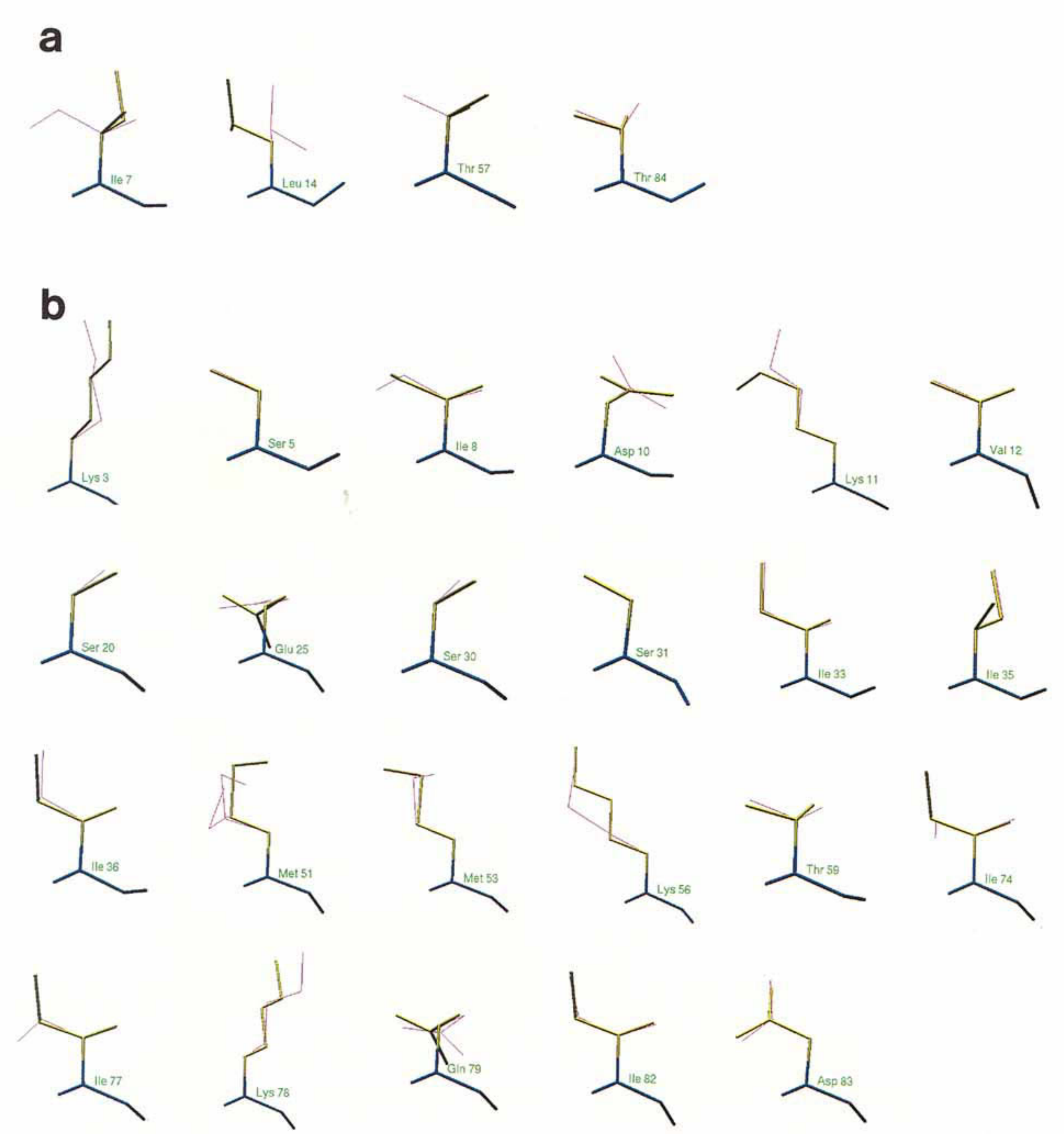

Fig. 4. (**a**) Nonconserved side chains of hpr with different conformations in the prediction and solution. Backbone movements and water molecules in the crystal solution appear to be the confounding factors. (**b**) Nonconserved side chains with similar energy and the correct conformation. The side chains shown have *B*-factors of low or moderate magnitude in the X-ray solution. Where a distal atom shows divergence, the X-ray data show a high *B*-factor for that atom alone in the side chain (Asp-10, Ile-74, Ile-77) indicating that the prediction is correct.

throughout the backbone which sum to at least 350 kcal.

DISCUSSION

Let us separate four components of the modeling by homology problem and focus on the central two of them. The components are (1) threading the target sequence onto a homologous structure, (2) nonconservative side-chain placement, (3) loop prediction, and (4) global refinement including the backbone.

Threading Was Accurately Accomplished in Each of the Targets

We find that manual threading techniques are sufficient to prepare a model with intermediate to high sequence identity homologies. Even if the

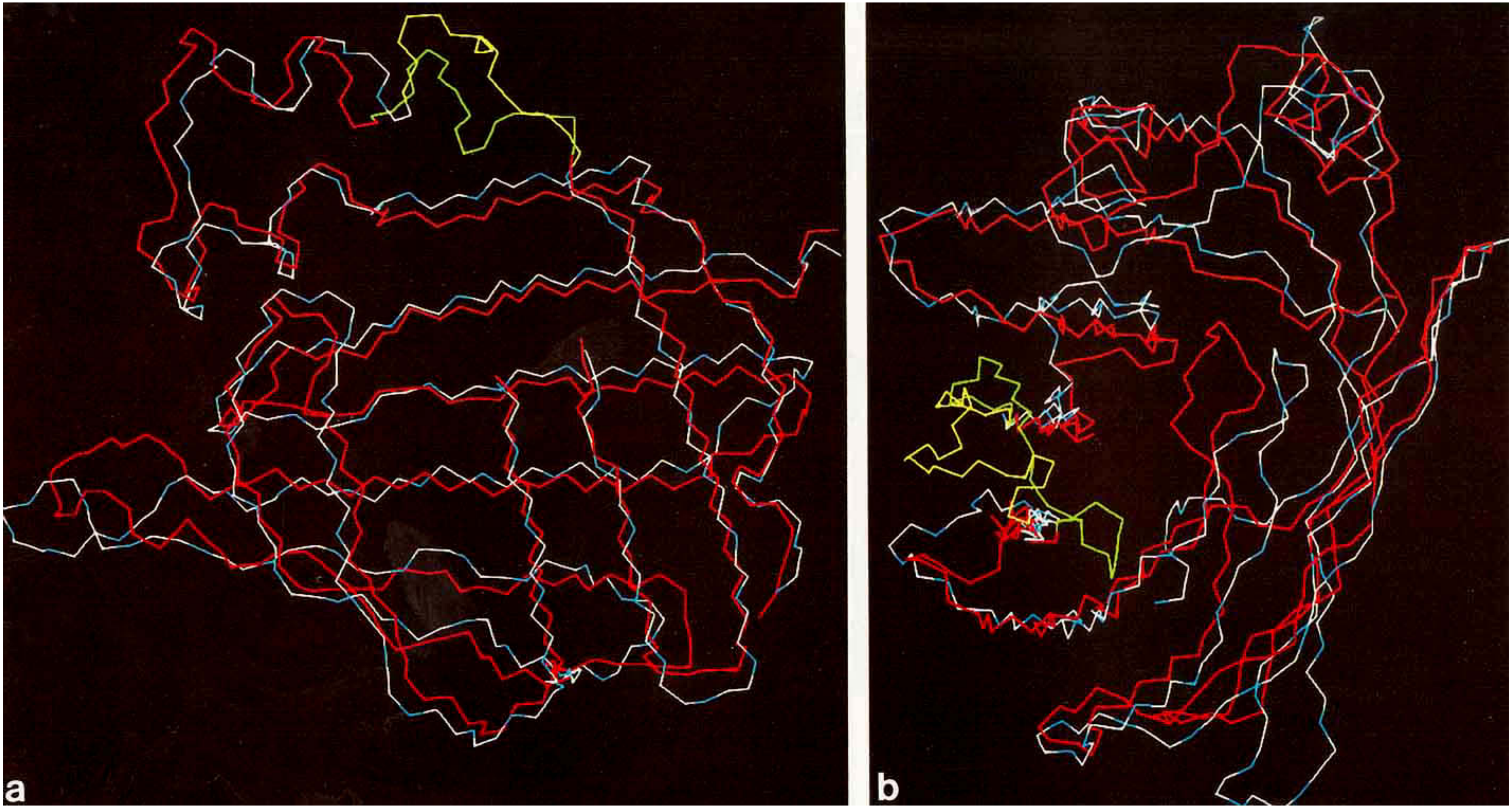

Fig. 5. Two loops modeled in cellular retinoic acid binding protein. Backbones of prediction (natural) and solution (red). Prediction loop is highlighted in yellow, solution in green. (**a**) Loop 1: the helix-turn-helix is displayed toward the top of the figure. The turn is visibly displaced to the left by the extra helical turn formed (green). (**b**) Loop 2: The conserved backbone movement shown at the center of the figure interacts with the loops.

TABLE III. Energy Values for Given Conformations of the Loops Modeled*

Target	Modeled	Energy term	Prediction	Near-native	Native
crabp	Loop 1 (res 34–39): α-helical solution	van der Waals	980.963	1186.009	40.078
	Loop 1 (res 34–39): α-helical solution	Hydrogen bond	34.341	28.654	−17.115
	Loop 2 (res 100–108): β-strand solution	van der Waals	264.382	6814.194	119.479
	Loop 2 (res 100–108): β-strand solution	Hydrogen bond	2.505	695.046	−17.642
antibody_ E53	H3 Loop (res 198–211)	van der Waals	−128.473	242.491	−297.447

*Only those terms shown exhibited significant differences between the conformations compared (loop 1 hydrogen bond energy is shown for comparison with loop 2 hydrogen bond). Near-native refers to the solution conformation of the loop in the setting of the model. Native refers to the solution conformation of the entire protein with the rigid body portion of the model applied. In all cases, the total energy of the near-native conformation was higher than the predicted [e.g., even though the hydrogen bonding energy of the near-native in crabp loop 1 (row 2) is lower, the much higher van der Waals energy (row 1) outweighs it].

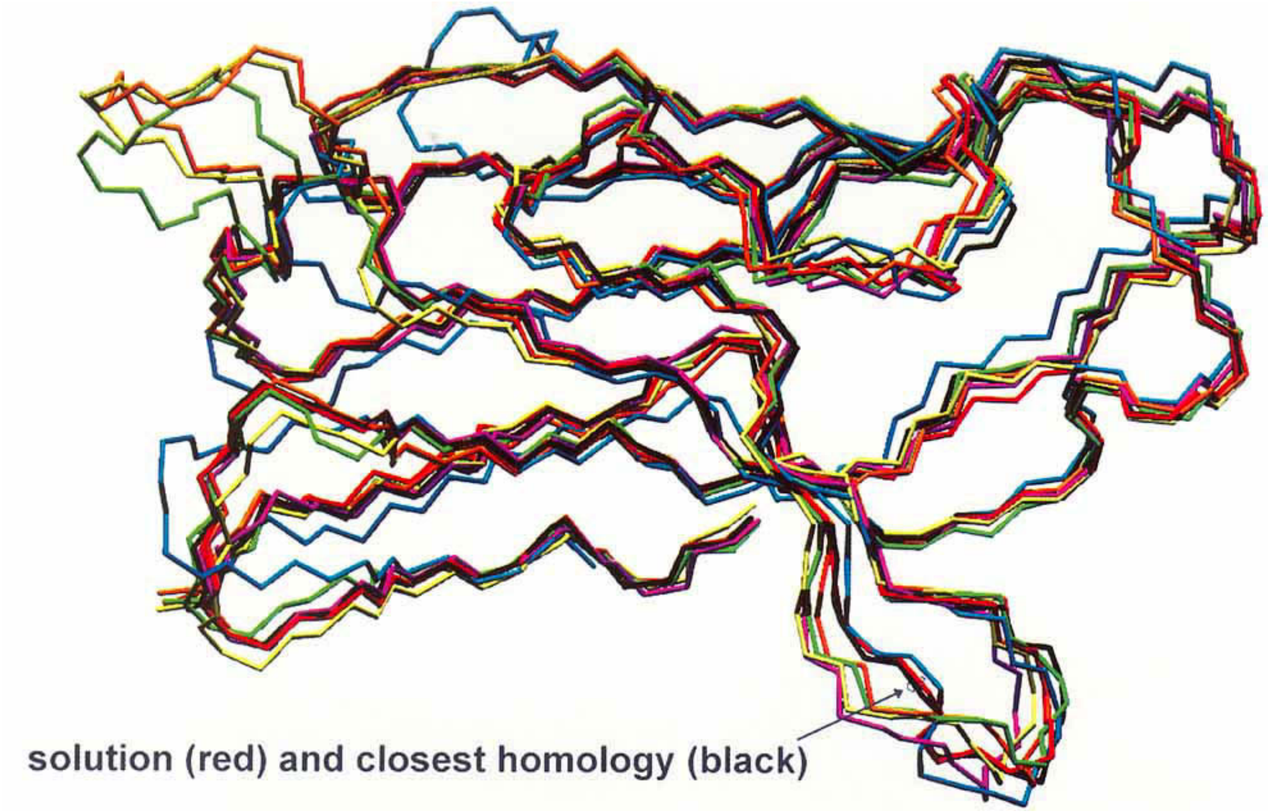

Fig. 6. Superimposition of the backbone of 6 candidate PDB light chains for the E5.2 antibody modeling target with the solution structure. The starting structure (6fab, traced in black) is the closest structure spatially to the solution (red). It was also the closest structure by pairwise sequence alignment (89% identity). All the light chains shown are of high sequence identity.

threading is completely accurate, which is increasingly less certain as one moves down the sequence identity scale to the moderate and low range, the rms distortions of the starting structure leave it outside the radius of convergence of local minimization procedures or Cartesian molecular dynamics.[31] The one low sequence identity test case we modeled accurately accomplished the residue assignment and yet was still clearly beyond any hope of improvement. This makes subsequent optimization pointless and reduces the entire homology modeling procedure to the residue assignment. Although it is clear that at the level of sequence identity where alignment is nontrivial further modeling does not bring the model backbone closer to the solution, the question then remains as to where the threshold level is at which further modeling makes sense.

The Nearest Identity Structure for Any Given Locale in the Protein Is a Better Starting Conformation Than an Average of Homologous Structures

All the CATPSP meeting targets had homology with more than one PDB structure. Using an average of homologous structures[20] as opposed to a single structure closest in sequence identity for a given stretch of the protein would almost certainly make the modeling task more difficult.[44] For one, an average structure would be useful only if the homologies were of equal evolutionary distance from the solution or if it were somehow possible to weight the average around the solution, both near impossibilities. Furthermore, even for small differences in sequence distance, the chances are small that the average structure will be closer than the nearest since the distribution is almost certainly uneven and statistics show accelerated rmsd in the intermediate range of identity of the PDB.[19] The light chain of the E5.2 antibody target illustrates this point (Fig. 6).

Side-Chain Placement Was Successful

The results of our side chain predictions are encouraging. In almost all of the individual side chains examined, the prediction is either correct or unverifiable due to the nature of the solution data. Thus, where the prediction of a side chain can be isolated from the task of predicting the backbone (i.e., at high sequence identity), ICM can accomplish the prediction. This means both that the energy function is able to recognize the near-native conformation and that BPMC sufficiently samples the phase space of nonconserved side chain conformations. In addition, since coupling between side chains was not a confounding factor, this prediction is another illustration that the combinatorial problem[6] of simultaneous side chain prediction is irrelevant.[1] In those cases when the side chain was predicted incorrectly, it was not because of insufficient sampling, but rather due to backbone displacements which make the near-native conformation of higher energy and thus unrecognizable.

If the Energy Function Is Reliable and the Sampling Is Sufficient, Loop Modeling Reveals a High Degree of Dependence on Global Backbone Refinement

In each of the loop modeling cases we attempted, backbone shifts and interactions adjacent to the loop being predicted made correct prediction impossible. All three near-native conformations exhibited higher energy in the fixed setting of the homologues structure and thus would not have been selected by our procedure. Transposition of the parameters of each model to the entire native structure reveals that the energy differences are largely due to van der Waals strain and, in one case, hydrogen bonding. Visual inspection reveals large adjacent backbone shifts in two of the loops (crabp) which may be responsible for the strain. Without these backbone shifts, which could not have been predicted with confidence, the native conformation results both in higher van der Waals and hydrogen bonding energy. Even small backbone shifts, as in the H3 loop of E5.2 antibody, made the energy of the near-native local conformation too high.

It does not appear, then, from any of these loop modeling cases, that the energy function utilized is unreliable. We still may address whether or not our procedure results in sufficient sampling to predict the largest of these loops (13 residues). Our procedure found a stable minimum from 4 different starting conformations. In work published elsewhere,[13] a designed loop of 8 residues in triosphosphate isomerase was correctly predicted. Although the TIM loop prediction represented the ideal case isolated modification on a known three-dimensional structure, the sampling afforded by our procedure appears to be sufficient.

What can we hope to accomplish with current techniques, and more specific to this work, what are the capabilities of ICM and the Biased Probability Monte Carlo sampling in modeling by homology? Our results suggest that the problem of sufficient sampling, which has been of primary concern to this point, can be addressed with the BPMC method. Furthermore, the discriminating function (energy) appears to be reliable. Therefore, problems purely of sampling (side-chain prediction, modeling of loops with very well-defined boundary and environmental conditions) may be solved. The limiting size of such problems remains to be defined. We have accomplished here the prediction of a maximum of 39 side chains and elsewhere of a loop 8 residues in length.[13] With these cases in mind, it seems likely that any pure sampling problems of subsets consisting of side chains only or loops of the range of sizes commonly found in currently solved structures may be accomplished with current computational power. If sufficient sampling is achievable, problems in which very limited discrete and identifiable backbone *shifts* complicate the prediction may be solved by the ICM method. The loop at residues 35–36 in the cellular retinoic acid binding protein is one such problem. While we have treated the substance of the protein not directly in the environment of the predicted loop as a single rigid body, this case points to the possibility of solution if that single rigid body is broken up into two or more movable pieces allowing axial and lateral displacements of entire blocks of secondary structure to accommodate, for example, the extension of a helix or β-strand.

The prediction of a loop in many of the situations presented by homology modeling, however, appears still to be beyond the reach of the technique. Interestingly, the adequacy of our energy function and sampling has uncovered the complexity of isolating the loop prediction component of modeling from global refinement. This isolation is necessitated by the capabilities of currently available computer power, but may have to be abandoned in pursuit of computationally elusive global prediction techniques.

REFERENCES

1. Eisenmenger, F., Argos, P., Abagyan, R.A., A method to configure protein side-chains from the main-chain trace in homology modeling. J. Mol. Biol. 231:849–860, 1993.

2. Dunbrack, R.L., Karplus, M. Backbone-dependent rotamer library for proteins: Application to side-chain prediction. J. *Mol. Biol. 230*:543–574, 1993.
3. Holm, L., Sander, C. Fast and simple Monte Carlo algorithm for side chain optimization in proteins: Application to model building by homology. Proteins 14:213–223, 1992.
4. Desmet, J., Maeyer de M., Hazes, B., Lasters, I. The dead-end elimination theorum and its use in protein side-chain positioning. Nature (London) 356:539–542, 1992.
5. Abagyan, R.A., Argos, P. Chemtracts Biochem. Prediction of protein side-chain conformation by packing optimization. Mol. Biol. 2:324–327, 1992.
6. Tuffery, P., Etchebest, C., Hazout, S., Lavery R. A new approach to the rapid determination of protein side chain conformations. J. Biomol. Struct. Dyn. 8:1267–1289, 1991.
7. Lee, C., Levitt, M. Prediction of protein side-chain conformation by packing optimization. Nature (London) 352: 448–451, 1991.
8. Lee, C., Subbiah, S. Prediction of protein side-chain conformation by packing optimization. J. Mol. Biol. 217:373–388, 1991.
9. Holm, L., Sander, C. Database algorithm for generating protein backbone and side-chain co-ordinates from a C-alpha trace. Application to model building and detection of co-ordinate errors. J. Mol. Biol. 218:183–194, 1991.
10. Schiffer, C.A., Caldwell, J.W., Kollman, P.A., Stroud, R.M. Prediction of homologous protein structures based on conformational searches and energetics. Proteins 8:30–43, 1990.
11. Summers, N.L., Karplus, M. Construction of side chains in homology modelling: Application to the C-terminal lope Rhizopuspepsin. J. Mol. Biol. 210:785–811, 1989.
12. 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.
13. 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.
14. 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.
15. Bowie, J.U., Eisenberg, D. Inverted protein structure prediction. Curr. Opinion Struct. Biol. 3:437–444, 1993.
16. Wodak, S.J., Rooman, M.J. Generating and testing protein folds. Curr. Opinion Struct. Biol. 3:247–259, 1993.
17. Jones D.T., Taylor W.R., Thornton J.M. A new approach to protein fold recognition. Nature (London) 358:86–89, 1992.
18. Abagyan, R.A., Frishman, D., Argos, P. Recognition of distantly related proteins through energy calculations. Proteins 19:132–140, 1994.
19. Hilbert, M., Bohm, G., Jaenicke, R. Structural relationships of homologous proteins as a fundamental principle in homology modeling. Proteins 17:138–151, 1993.
20. Blundell, T., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D.A., Sibanda, B.L., Sutcliffe, M. Knowledge-based protein modelling and design. Eur. J. Biochem. 172:513–520, 1988.
21. Summers, N.L., Karplus, M. Modeling of side-chains, loops, and insertions in proteins. Methods Enzymol. 202: 156–205, 1991.
22. Schrauber, H., Eisenhaber, F. Argos, P. Rotamers: To be or not to be? An analysis of amino acid side-chain conformations in globular proteins. J. Mol. Biol. 230:592–612, 1993.
23. Martin, A.C.R., Cheetham, J.C., Rees, A.R. Modeling antibody hypervariable loops: A combined algorithm. Proc. Natl. Acad. Sci. U.S.A. 86:9268–9272, 1989.
24. Chothia, C., Lesk, A.M., Levitt, M., Amit, A.G., Mariuzza, R.A., Phillips, S.E., Poljak, R.J. The predicted structure of immunoglobulin D1.3 and its comparison with crystal structure. Science 233:755–758, 1986.
25. Bruccoleri, R.E., Karplus M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 26:137–168, 1987.
26. Bruccoleri, R.E. Application of systematic conformational search to protein modeling. Mol. Simulation 10:151–174, 1993.
27. Shenkin, P., Yarmush, D.L., Fine, R.M., Wang, H., Levinthal, C. Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ring-like structures. Biopolymers 26:2035–2085, 1987.
28. Mas, M.T., Smith, K.C., Yarmush D.L., Aisaka, K., Fine R. M. Modeling the anti-CEA combining site by homology and conformational search. Proteins 14:483–498, 1993.
29. Higo, J., Collura, V., Garnier, J. Development of an extended simulated annealing method: Application to the modeling of complementary determining regions of immunoglobulins. Biopolymers 32:33–43, 1992.
30. Collura, V., Higo, J., Garnier, J. Modeling of protein loops by simulated annealing. Protein Sci. 2:1502–1510, 1993.
31. Abagyan, R.A., Totrov, M.M., Kuznetsov, D.A. 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.
32. Abagyan, R.A. Towards protein folding by global energy optimization. FEBS Lett. 1:17–22, 1993.
33. 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.
34. Totrov, M.M., Abagyan, R.A. Detailed ab initio prediction of lysozyme-antibody complex with 1.6 Å accuracy. Nature Struct. Biol. 1:259–263, 1994.
35. Higgins D.G., Bleasby A.J., Fuchs R. CLUSTAL V: Improved software for multiple sequence alignment. CABIOS 8:189–191, 1992.
36. Thompson, J.D., Higgins, D. G., Gibson, T. Improved sensitivity of profile searches through the use of sequence weights and gap excision. J. Comput. Appl. Biosci. 10:19–30, 1994.
37. Thompson, J.D., Higgins, D.G., Gibson, T. Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. J. Acids Res. 22: 4673–4680, 1994.
38. Momany, F.A., McGuire, R.F., Burgess, A.W., Scheraga, H.A. Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occuring amino acids. J. Phys. Chem. 79:2361–2381, 1975.
39. Nemethy, G., Pottle, M.S., Scheraga, H.A. Energy parameters in polypeptides. IX. Udating of geometric parameters, nonbonded interactions and hydrogen bond interactions for the naturally occuring amino acids. J. Phys. Chem. 87:1883–1887, 1983.
40. Nemethy, G., Gibson, K.D., Palmer, K.A., Yoon, C.N., Paterlini, G., Zagari, A., Rumsey, S., Scheraga, H.A. Energy parameters in polypeptides. X. Improved geometric parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. J. Phys. Chem. 96:6472–6484, 1992.
41. Wesson, L., Eisenberg, D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Protein Sci. 1:227–235, 1992.
42. Abagyan, R., and Totrov, M. 1995 Unpublished.
43. Bassolino-Klimas, D., Bruccoleri, R.E., Subramaniam, S. Modeling the antigen combining site of an anti-dinitrophenyl antibody, AN02. Protein Sci. 1:1465–1476, 1992.
44. Sali, A., Blundell, T. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234:779–815, 1993.
45. Palmer K.A., Scheraga, H.A. Standard-geometry chains fitted to X-ray derived structures: Validation of the rigid-geometry approximation. I. Chain closure through a limited search of "loop" conformations. J. Comp. Chem. 12: 505–526, 1991.
46. Palmer K.A., Scheraga, H.A. Standard-geometry chains fitted to X-ray derived structures: Validation of the rigid-geometry approximation. II Systematic searches for shout loops in proteins: Applications to bovine pancreatic ribonuclase A and human lysozyme. J. Comp. Chem. 13:329–350, 1992.