PROTEINS: Structure, Function, and Genetics 23:318–326 (1995)

Evaluation of Comparative Protein Modeling by MODELLER

Andrej Šali,[1,3] Liz Potterton,[2] Feng Yuan,[3] Herman van Vlijmen,[3] and Martin Karplus[3]
[1]The Rockefeller University, New York, NY 10021, [2]Molecular Simulations Inc., Department of Chemistry, University of York, York Y01 5DD, UK, and [3]Department of Chemistry, Harvard University, Cambridge, Massachusetts 02138

ABSTRACT **We evaluate 3D models of human nucleoside diphosphate kinase, mouse cellular retinoic acid binding protein I, and human eosinophil neurotoxin that were calculated by MODELLER, a program for comparative protein modeling by satisfaction of spatial restraints. The models have good stereochemistry and are at least as similar to the crystallographic structures as the closest template structures. The largest errors occur in the regions that were not aligned correctly or where the template structures are not similar to the correct structure. These regions correspond predominantly to exposed loops, insertions of any length, and nonconserved side chains. When a template structure with more than 40% sequence identity to the target protein is available, the model is likely to have about 90% of the mainchain atoms modeled with an rms deviation from the X-ray structure of ≈ 1 Å, in large part because the templates are likely to be that similar to the X-ray structure of the target. This rms deviation is comparable to the overall differences between refined NMR and X-ray crystallography structures of the same protein.**

Key words: evaluation, comparative protein modeling

INTRODUCTION

Protein modelers were challenged to model sequences without available three-dimensional (3D) structures and to submit them to the first "Meeting on Critical Assessment of Techniques for Protein Structure Prediction" in Asilomar in December of 1994.* At the same time, the 3D structures were being determined by X-ray crystallography or nuclear magnetic resonance (NMR) methods. Because these structures were only released at the meeting, it was possible to test the modeling methods objectively.

We submitted homology-derived models of three proteins: Human nucleoside diphosphate kinase (NM23H2), mouse cellular retinoic acid binding protein I (CRABP1), and human eosinophil neurotoxin (EDN). All three structures have been determined by X-ray crystallography: NM23H2 at 2.8 Å resolution and R-factor of 24% (R. Williams, in preparation), CRABP1 at 2.9 Å resolution and R-factor of 25%,[1] and EDN at 1.8 Å resolution and R-factor of 17% (S.C. Mosimann, D. Newton, R. Youle, and M.N.G. James, in preparation). These three sequences were chosen because they span a range of difficulty from easy (NM23H2, based on 77% sequence identity with templates), and medium (CRABP1, 41%), to difficult (EDN, 33%).

Our approach to comparative protein modeling is based on satisfaction of spatial restraints and is implemented in program MODELLER.[2–7,†] This program can be used in all stages of typical comparative modeling: Finding suitable template structures in the Brookhaven Protein Databank (PDB),[8] aligning them with the sequence to be modeled, calculating the 3D model, and evaluating the model. An application of MODELLER to site-directed mutagenesis study of heparin binding by mouse mast cell proteases has been described.[9,39] Comparative protein modeling was recently reviewed.[4,10–12]

In this paper, we briefly describe modeling of the three proteins and then concentrate on evaluation of the models.

METHODS

This section outlines all the stages in calculation of the 3D models. With the exception of tree clustering,[13] all the procedures were carried out by MOD-

*The organizing committee was composed of John Moult, Jan Pedersen, Krzysztof Fidelis, Rod Balhorn, Richard Judson, and Walt Stevens.

†MODELLER is available by anonymous FTP from guitar.rockefeller.edu:pub/modeller and also as part of QUANTA (MSI, Burlington, MA, USA; e-mail jcollins@msi.com).

Abbreviations: CRABP1, mouse cellular retinoic acid binding protein I; NM23H2, human nucleoside diphosphate kinase; NMR, nuclear magnetic resonance; EDN, human eosinophil neurotoxin; PDB, Brookhaven Protein Data Bank; rms, root-mean-square; 2D, two-dimensional; 3D, three-dimensional.

Received April 3, 1995; revision accepted June 9, 1995.
Address reprint requests to A. Šali, The Rockefeller University, Box 270, 1230 York Avenue, New York, NY 10021.

ELLER-2 automatically without any subjective decisions.

Finding Structures Related to the Target Sequence

Proteins that have known 3D structure and are similar to the sequences being modeled (target sequences) had to be identified. This was achieved by searching a set of sequences representative of the whole PDB (September 1, 1994), using the SEQUENCE_SEARCH command of MODELLER.[7] The representative set of proteins includes 627 structures whose sequence identity is less than 30% to any other structure in the set. Each representative sequence was aligned with the target sequence by a global dynamic programming algorithm.[14] The significance score of each alignment was expressed in terms of standard deviations from the mean score of 100 optimal alignments of random sequences; these random sequences were obtained from the original two sequences by randomly shuffling their amino acid residues. If the significance score of an alignment was larger than 4.0, the representative sequence and all its structurally defined homologs with sequence identity higher than 30% were considered as potential templates for modeling.

All sequence alignments relied on the 20 by 20 residue–residue weight matrix as1.mat that was derived from a database of protein structure alignments[5] and the gap initiation and extension penalties that were determined by optimization of the significance scores for pairs of remotely related structures (F. Yuan and A. Šali, unpublished). This procedure has no difficulty in finding related sequences in the domain suitable for comparative protein modeling, that is for sequence identities above 25%.

Aligning All Structures With the Target Sequence

An alignment of all the structures was prepared by multiple least-squares superposition,[15] as implemented in the MALIGN3D command.[7] The multiple structural alignment was then aligned as one block with the corresponding target sequence, using the dynamic programming method implemented in the MALIGN command.[7]

Selecting the Templates

Matrices of pairwise sequence identities were calculated from the family alignments and employed to construct "evolutionary" trees for the three families, relying on the KITSCH program from the PHYLIP package.[13] All significantly different structures in the cluster that contained the target sequence were used as templates in the subsequent model building.

Model Building

The alignments and the lists of templates were used, without manual intervention, to calculate 3D models for the three sequences containing all mainchain and side chain heavy atoms. First, MODELLER derived many distance and dihedral angle restraints on the target sequence from its alignment with template 3D structures.[4] Second, the spatial restraints and energy terms enforcing proper stereochemistry[16] were combined into an objective function.[4] Third, the models were obtained by optimizing the objective function in Cartesian space.[4] This optimization was carried out by the use of the variable target function method employing methods of conjugate gradients and molecular dynamics with simulated annealing. This procedure optimizes the positions of all atoms at the same time. Only five models were derived for each sequence by varying the initial structure because no models with significantly lower value of the objective function are generally obtained if a larger number of models is calculated. The representative model was that which had the lowest value of the objective function.

Evaluating the Models

Before the models were submitted, they were evaluated for self-consistency. The models had to satisfy most restraints used to calculate them, especially the stereochemical restraints. These tests were done by the MODELLER's ENERGY command,[7] the PROCHECK program,[17] and by the options in the QUANTA Protein Health module (MSI, Burlington, MA). Additional evaluation was done by 3D profile programs PROSAII[18] and 3DPROFILE,[19] and by comparison with the corresponding X-ray structures.

Modeling Loop 115–123 in EDN

An alternative model for the surface loop of residues 115–123 in EDN was calculated by the use of a method relying on a PDB database search and energy minimization (H. van Vlijmen and M. Karplus, in preparation). Residues 112–114 and 124–126 were defined as the stem residues. Scanning a representative database of 173 proteins, 1100 13-residue segments whose N- and C-termini superposed well with the stem residues were identified.[20] The resulting template loops were superimposed on the EDN model using the stem residues. All side chains, except for Pro-120 and Pro-121, were represented by C_β atoms only. Using CHARMM,[16] the poly-(Ala,Pro) loop was minimized in the field of all remaining atoms of the modeled protein. The 171 lowest energy loops were further minimized, sidechains were added according to the rotamer libraries of Ponder and Richards[21] and Dunbrack and Karplus,[22] and the resulting template loops were again minimized until convergence. The template loop with the low-

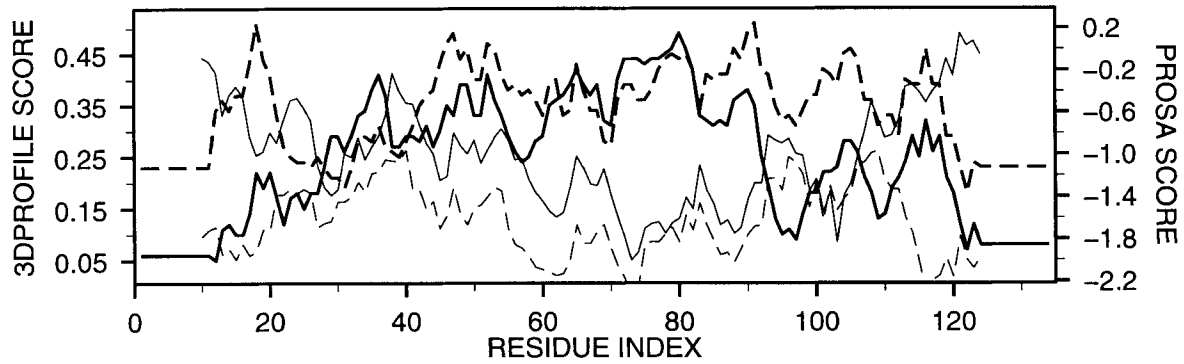

Fig. 1. 3D profile tests of the EDN model and X-ray structure. 3DPROFILE[19] profiles are shown in thick lines. PROSAII[18] profiles are shown in thin lines. Model (continuous lines). X-ray structure (dashed lines). Scores larger and smaller than 0 are deemed acceptable for 3DPROFILE and PROSAII, respectively.

est total energy was selected as an alternative to the MODELLER model.

RESULTS

Stereochemistry of the Models

The models satisfy all stereochemical restraints used to calculate them. For example, for the NM23H2 model, the rms deviation from the ideal values for bond lengths, angles, and improper dihedral angles is 0.005 Å, 2.2°, and 1.3°, respectively. Also, the model passed all criteria implemented in PROCHECK[17] to the same degree as the X-ray structures refined at 1.5 Å resolution. These criteria are Ramachandran plot, peptide bonds planarity, C_α tetrahedral distortion, nonbonded interactions, hydrogen bond energies, and closeness of side chain dihedral angles to ideal values. For example, for the NM23H2 model, 97% of residues reside in the most favored areas of the Ramachandran plot (90% for structures solved at $\leq$ 2 Å resolution) and the overall *G*-factor measuring stereochemical quality is 0.3 (−0.5 to 0.3 for structures solved at 1.5 Å resolution).

3D Profiles of the Models

A more stringent test of the overall fold and side chain packing of the models may be provided by the 3D profile programs.[19,23] These methods plot, in a moving window along the sequence, a score or energy that is correlated with the accuracy of the model in the corresponding region and depends only on the sequence and structure of the model. Such profiles from the 3DPROFILE[19] and PROSAII[18] programs are shown for the least accurate model, that of EDN, in Figure 1. The profiles for the model are compared with those for the highly refined X-ray structure of EDN. The X-ray structure generally scores better than the model. This opens the possibility that the accuracy of the model could be improved by adding the 3D profile terms to the objective function optimized by MODELLER, although it is not clear how many incorrect models also have better scores than the current MODELLER model. Unfortunately, none of the errors in the three modeling cases could have been identified by the 3D profiles because even the incorrect regions in the models have reasonably good scores, i.e., the 3DPROFILE scores are larger than 0 and the PROSAII scores are smaller than 0.

NM23H2 Model

NM23H2 is a single domain protein consisting of a central 4-stranded antiparallel β-sheet surrounded by 8 α-helices. The crystallographic structure has been determined for a hexamer ring of identical subunits (R. Williams, in preparation). Of the 10 PDB coordinate sets with significant sequence similarity to NM23H2 (see Methods), four were used to model NM23H2: Nucleoside diphosphate kinase from *Drosophila melanogaster* (with the PDB code of 1NDL and 77% sequence identity to NM23H2) and three forms of nucleoside diphosphate kinase from *Dictyostelium discoideum* (1NDC, 1NDP, 1NDK; 60%).

The most important evaluation of a model is its comparison to the corresponding X-ray structure. The atomic positions of C_α, mainchain, and all atoms are compared using two interdependent criteria: The number of atoms that superpose within 3.5 Å from each other (equivalent atoms) and the rms difference between those atoms (Table I).

For the NM23H2 model, all but one C_α atom superpose within 3.5 Å of the X-ray structure; rms difference is 0.41 Å. 95% of all atoms superpose with an rms of 0.93 Å. Approximately 76, 65, and 58% χ_1, χ_2, and χ_3 side chain rotamers, respectively, are modeled in accordance with the X-ray structure. Part of the model is compared with the X-ray structure in Figure 2. To put these numbers into perspective, they are contrasted with the differences between two independently refined subunits in the hexamer of the NM23H2 X-ray structure (Table I):

TABLE I. Comparison of Models With Templates and Corresponding X-Ray Structures

Structures	C_α		Main chain		All		Dihedral angle classes‡				
	Atoms*	rms†	Atoms	rms	Atoms	rms	(Φ,Ψ)	χ_1	χ_2	χ_3	χ_4
NM23H2-U$^{\Sigma}$–MODEL	148/152	0.41/147	588/607	0.43/587	1179/1218	0.93/1123	136/11	96/31	64/35	28/20	7/15
NM23H2-U–NM23H2-R	148/147	0.50/147	588/586	0.55/586	1179/1178	0.92/1152	133/12	98/28	70/29	23/25	11/11
MODEL–1NDL	152/152	0.18/152	607/606	0.18/606	1218/1203	0.56/1129	148/2	114/17	80/23	35/15	9/13
NM23H2-U–1NDL	148/152	0.44/147	588/607	0.45/587	1179/1203	0.87/1075	136/11	96/31	64/35	28/20	7/15
CRABP1–MODEL	136/137	1.31/122	543/547	1.34/485	1087/1094	1.57/885	110/25	76/42	61/25	25/15	7/10
CARBP1–2HMB	136/131	1.34/122	543/487	1.38/477	1087/1031	1.53/775	106/15	52/56	41/37	20/18	6/12
CRABP1–1LIF	136/131	1.33/120	543/524	1.33/470	1087/1017	1.45/743	104/15	54/52	36/41	15/21	9/9
MODEL–2HMB	137/131	0.49/131	547/487	0.59/521	1094/1031	0.84/861	116/13	74/41	43/41	14/25	8/9
EDN–MODEL	134/134	1.17/90	535/535	1.25/366	1081/1081	1.43/650	92/40	78/47	53/46	23/27	9/11
EDN–7RSA	134/124	1.14/109	535/495	1.17/428	1081/951	1.25/668	88/20	61/40	37/41	12/22	6/10

*The numbers of atoms of the indicated type in the first and second structure, respectively.

†The number of atoms within 3.5 Å from each other is indicated in the denominator. The rms difference between them is shown in Å in the nominator.

‡The numbers of identical and different dihedral angle classes are shown for all aligned amino acid positions. For identical proteins, all residues are aligned. For different proteins, only those residues that have C_α atoms within 3.5 Å from each other are aligned. The Ramachandran plot is divided into six areas corresponding to the six (Φ,Ψ) classes: right-handed α-helix, idealized β-strand, polyproline conformation, left-handed α-helix with Gly with positive Φ.[4] The side chain dihedral angle classes correspond to rotamers, generally 3 per dihedral angle.[4]

$^{\Sigma}$NM23H2-U and NM23H2-R are the U and R subunits, respectively, in the crystallographically determined structures of the NM23H2 hexamer.

The model is slightly more similar to some of the crystallographic subunits than some of the crystallographic subunits are to each other. While the differences among the crystallographic subunits may decrease with higher resolution and refinement, this result indicates that a homology model derived from templates with 70–80% sequence identity is as accurate as a medium-resolution X-ray structure. In large part, this is a consequence of the corresponding similarity between the template and target structures (Table I).

CRABP1 Model

CRABP1 is a single domain protein composed of interacting α-helices packed at the edge of two orthogonal, 4- and 6-stranded antiparallel β-sheets.[1] Of the 15 PDB coordinate sets with significant sequence similarity to CRABP1 (see Methods), two were used to model CRABP1: fatty acid binding protein (2HMB; 41%) and mouse adipocyte lipid-binding protein (1LIF; 38%).

For the CRABP1 model, 90% of C_α atoms superpose within 3.5Å of their counterparts in the X-ray structure; the rms error is 1.31 Å (Fig. 3). The errors are concentrated in four regions (Fig. 4): (1) At the end of the helix at residue 38, where there is an insertion of two residues relative to the two templates (several of the CRABP1 models, but not the representative model, had the 2 residue insertion at residue 38 modeled correctly as a continuation of the helix). (2) At the tip of the β-hairpin at residue 46, where 5 CARBP1 residues move as a rigid body relative to the templates. (3) At the tip of the β-hairpin at residue 76, where the hairpin is twisted relative to the two template structures. (4) At the tip of the loop at residue 104, where there is a 4 residue insertion relative to the two templates and the sequence is modeled as a β-hairpin whereas the X-ray structure has a wider loop. Loops at residues 38, 76, and 104 are close in space. It appears that segments at residues 76 and 104 move, relative to the two templates, to fill the space emptied by a shift of the helix at residue 38. Thus, the largest four errors result from the structural differences between the templates and the X-ray structure or from the lack of equivalent residues in the template structures, not from incorrect alignment. The alignment is nearly perfect despite the fact that it was derived automatically on the basis of sequence alone. Judging by the structural superposition of the templates and the X-ray structure of CRABP1, only a single residue neighboring a gap in the alignment should be aligned with another residue and only 4 single positions in loop regions should be aligned with a gap instead of a residue.

It is clear from Table I that the CRABP1 model is only slightly closer, in terms of overall rms values, to the crystallographic structure than the template structures. In other words, the number of equivalent positions and their rms values are similar for the comparison between the model and its X-ray structure and for the comparison between the CRABP1 X-ray structure and template structures. However, we note that the homology modeling methods some-

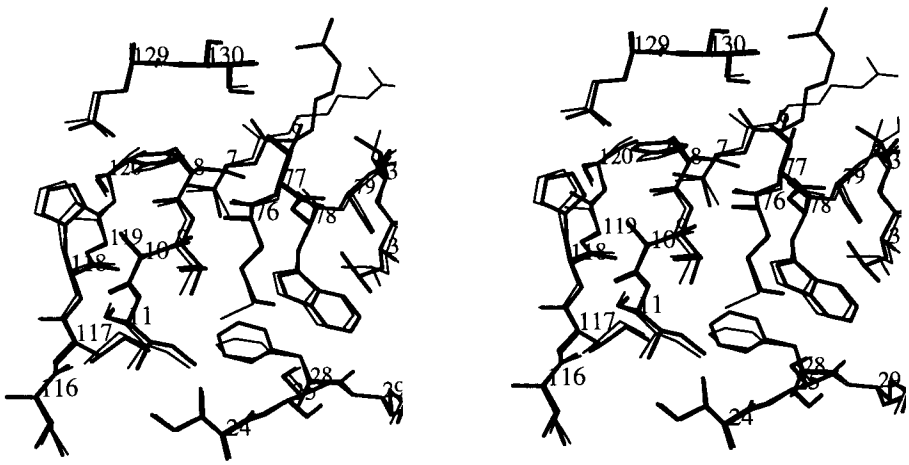

Fig. 2. Comparison of the NM23H2 model (thick lines) and the U subunit of the NM23H2 X-ray structure (thin lines). Part of the core region is shown. C_α atoms are labeled. All plots of protein structures were prepared by MOLSCRIPT.[38]

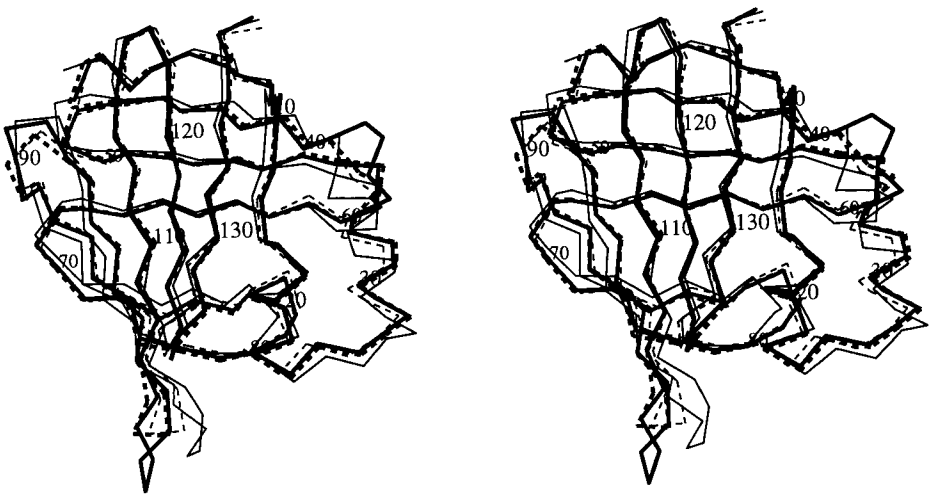

Fig. 3. Modeling of CRABP1. Superposition of C_α atoms of the CRABP1 model (thick continuous line), CRABP1 X-ray structure (thin continuous line), and the two templates, 2HMB (thick dashed line), and 1LIF (thin dashed line). Every tenth C_α atom in the CRABP1 model is labeled.

times produce models that have rms errors larger than the rms difference between the template structure and the X-ray structure of the sequence being modeled. Examples are listed in references[24] and in Mosimann et al. (this issue). Thus, it appears that it is not trivial to construct a 3D model with both good stereochemistry and with an rms deviation from the correct structure that is at least as small as that of the template's, i.e., for the target sequence, most ways of relaxing the template coordinates to improve the stereochemistry of the model increase the rms difference from the correct target structure.

Variability Among Models Correlates With Error—CRABP1

For a given sequence, an ensemble of slightly different models is always calculated, similarly to the refinement relying on NMR-derived restraints. Variability among the models reflects either the lack of strong restraints on a certain region or various alternatives offered by the structural differences among the templates. In either case, the variability in the models was expected to correlate with the error in the model. This expectation is borne out in Figure 4, which compares the variability among 15 CRABP1 models with positional differences between the equivalent C_α atoms in the CRABP1 model and X-ray structure. While one could predict a priori that the regions most likely to be in error are the loops, not all loops are highly variable in the models, and not all highly variable regions are loops. Thus, the variability plots can provide a more direct and quantitative information about the errors in a particular region, e.g., notice the correlation

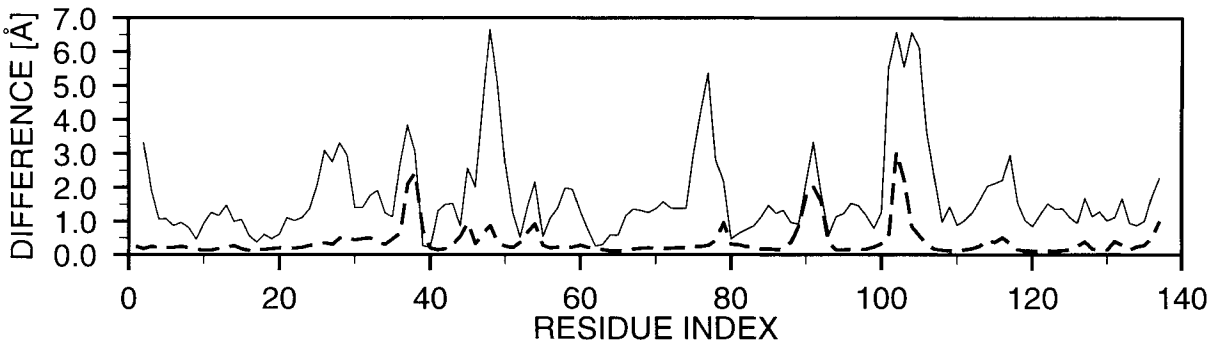

Fig. 4. Variability among the models is correlated with error. Fifteen CRABP1 models were superposed using their C_α atoms; the rms deviation of C_α atoms at each position is shown by the dashed line. Distance between equivalent C_α atoms in the optimally superposed representative model and the X-ray structure is shown by the continuous line.

between the degree of variability and the magnitude of error in Figure 4. Unfortunately, while all the variable regions correspond to maxima in error, not all regions in error are pointed out by the variability among the models. Also, the error is always larger than the variability. Thus, the variability imposes only a lower bound on error.

Alignment Errors—EDN

EDN is a ribonuclease with 3 α-helices and 2 three-stranded antiparallel β-sheets arranged in a single domain (S.C. Mosimann, D. Newton, R. Youle, and M.N.G. James, in preparation). Of the 15 PDB coordinate sets with significant sequence similarity to EDN (see Methods), only ribonuclease A (7RSA; 33%) was used to model EDN.

The source of the largest error in the EDN model was an incorrect alignment of its sequence with the template structure, 7RSA. The automated procedure attempted to minimize the number of gaps and thus shifted the N-terminal 21 residues of EDN by six positions relative to their correct equivalents in 7RSA (Fig. 5). The reason for this problem was the high gap-penalty which prevented the correct insertion of 9 7RSA residues that correspond to the loop following the first helix. While 3D profiles do not identify the model in this region as incorrect (Fig. 1), the alignment error could have been corrected by manual editing because lower gap penalties result in an almost correct alignment and because the active site His-15 residue was misaligned by the program.

Loop Modeling—EDN

The crystallographically determined EDN loop spanning residues 115–123 is compared to the representative MODELLER model, a loop obtained by a database search and energy minimization ("database model"; see Methods), and the equivalent loop in the template 7RSA structure, which has 9 residues fewer than the EDN loop (Fig. 6). Neither of the models is close to the X-ray structure, although the database model overlaps with it at positions 121–123. The MODELLER loop is too extended in comparison with the crystallographic structure. Some of the MODELLER models occupy a volume similar to that in the X-ray structure, but none has the correct conformation.

The best possible template that could have been obtained from the database had a main chain distance rms[25] of 1.17 Å compared to the crystallographic structure. The best possible template that was selected by the stem search had a distance rms of 1.23 Å. This loop was among the loops with the lowest energy, but not the very lowest one. With hindsight, the choice of stem residues at the N-terminus appears suboptimal because there is a distance of 4.1 Å between the last stem residue (114) in the template and the X-ray structure, which may account for the large error at the N-terminus in the database model. It would be better if residues 111–113 were used as a stem.

DISCUSSION

Homology models for NM23H2, CRABP1, and EDN were constructed automatically. The models were based on 77, 41, and 33% sequence identities with template structures, respectively. No significant improvement of the NM23H2 and CRABP1 models would have been achieved if subjective interventions had been made. On the other hand, a large error in the alignment of the EDN sequence with its template could have been prevented by examining and editing the alignment manually.

The models have good stereochemistry and are close to the template structures. The errors can be divided into four categories: (1) Errors in side chain packing. (2) Distortions or shifts of a region that is aligned correctly with the templates (e.g., loops, helices, strands). (3) Distortions or shifts of a region

```
                10        20        30        40        50        60
EDN   ---KPPQFTWAQWFETQHINMTSQQCTNAMQVINNYQRRCKNQNTFLLTTFANVVNVCGNPNMTCPSN
                                                                      ||
7RSA  KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHESLADVQAVCSQKNVAC-KN
          aaaaaaaaaaa      aaaaaaaaaaaaaa   bbbbbbb aaaaaaaaa

          70        80        90        100       110       120       130
EDN   KTRKNCHHSGSQVPLIHCNLTTPSPQNISNCRYAQTPANMFYIVACDNRDQRRDPPQYPVVPVHLDRII
                                                                ||||||||||
7RSA  -GQTNCYQSYSTMSITDCRETGSS--KYPNCAYKTTQANKHIIVACEGN---------PYVPVHFDASV
          bbbbb    bbbbbbbbbb     aaaaabbbbbbbbbbbbbbbbb          bbbbbbbbbbb
```

Fig. 5. Alignment of EDN and 7RSA. Automatically derived sequence alignment between EDN and 7RSA that was used for EDN modeling is shown. The black lines shows correct equivalences, that is residues whose C_α atoms are within 5 Å of each other in the optimal least-squares superposition of the X-ray structures of EDN and 7RSA. The bottom line indicates helices (a) and strands (b), as assigned in the EDN structure by program DSSP.[37]

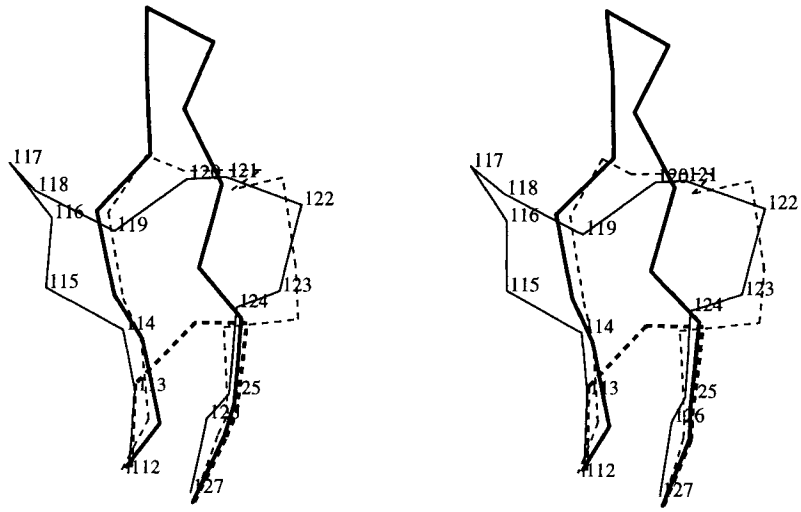

Fig. 6. Loop 112–127 in EDN. Stereo plot of the C_α trace of the 112–127 loop is shown for the EDN MODELLER model (thick continuous line), EDN database model (thin dashed line), EDN X-ray structure (continuous thin line), and the 7RSA template (residues 111–117; thick dashed line). The MODELLER loop comes from the representative model that was picked on the basis of the optimal *overall* objective function value, C_α atoms in EDN are labeled.

that does not have an equivalent segment in any of the templates (e.g., inserted loops). (4) Distortions or shifts of a region that is aligned incorrectly with the templates (e.g., loops and larger segments with low sequence identity to the templates). To address errors 1–3, a better potential function and possibly a better optimizer are needed. The potential function should guide the model away from the templates in the direction towards the correct structure. An addition of atomic or residue based potentials of mean force to the MODELLER objective function might be one way of achieving this goal. However, no present force field or potential of mean force can generally produce a model with a mainchain rms difference from the X-ray structure smaller than about 1 Å, even when the starting conformation is the X-ray structure itself. For example, molecular dynamics simulations in solvent generally have a main chain rms deviation of about 1 Å and the most detailed lattice folding simulations result in models with an rms > 2 Å.[26] Since most of the atoms in two homologs with at least 40% sequence identity usually superpose with a main chain rms difference of about 1 Å, it is currently better to aim to reproduce the template structures as closely as possible rather than to venture away from the templates in the search for a better model.

Errors 2–4 are relatively infrequent when sequences with more than 40% sequence identity with the templates are modeled. For example, in such a case, approximately 90% of the main chain atoms are likely to be modeled with an rms error of about 1 Å (Table I).

Below 40% sequence identity, misalignments and insertions in the target sequence become the major problems. It appears that insertions longer than about 8 residues cannot be modeled accurately at this time, even when the alignment of the stem regions delimiting the insertion is correct. Most of the insertions shorter than 8 residues also cannot be modeled successfully, primarily because the alignment of the inserted and neighboring residues is frequently incorrect. For example, there were 7 insertions in the CRABP1 and EDN models shorter than 8 residues, of which only 3 were aligned correctly, and only one of these was modeled correctly (loop at position 64 in EDN) and one was modeled correctly in some of the multiple models (loop at position 38 in CRABP1). If the length of an insertion can be extended enough to make the alignment of the delimiting stem regions reliable, but not too much so that less than 8 residues are inserted, the insertions can frequently be modeled successfully.[27–29] In general, it can be expected that about 20% of residues will be misaligned, and consequently incorrectly modeled when the target has 30% identity to the template.[30] Thus, the EDN–7RSA alignment errors (25 residues out of 134) are representative of comparative modeling in this range of sequence identity. To reduce the errors in the model stemming from the alignment errors, iterative changes in the alignment during the calculation of the model, perhaps similar to the threading techniques,[31,32] are needed.

To put the errors into perspective, we list the differences among experimentally determined structures. The 1 Å accuracy of mainchain atom positions corresponds to X-ray structures defined at a resolution of about 2.5 Å and with an R-factor of about 25%,[33] as well as to NMR structures determined from 10 interproton distance restraints per residue.[34,35] Similarly, differences between the highly refined X-ray and NMR structures of the same protein also tend to be about 1 Å.[34] Changes in the environment (e.g., crystal packing, solvent, ligands) can also have an effect on the structure on the order of 1 Å or even larger.[36] Overall, homology modeling based on templates with more than 40% sequence identity is almost as good (Table I), simply because the homologs at this level of similarity are likely to be as similar to each other as are the structures for the same protein determined by different experimental techniques under different conditions. The caveat in modeling, however, is that some regions, mainly loops and side chains, have larger errors. However, such loops also tend to differ more in different forms of the same protein. Although such regions may have an important function, many applications in biology do not require high resolution structures. For example, some binding sites may be located with the aid of low resolution models.[9,39]

Even though comparative modeling needs significant improvements, it is already a useful technique that can be employed by a nonexpert to address many practical problems. With the increase in the number of protein sequences and in the fraction of all folds that are known, comparative modeling will be even more useful in the future.

ACKNOWLEDGMENTS

We are grateful to crystallographers R. Williams, G. Kleywegt, A. Jones, S.C. Mosimann, D. Newton, R. Youle, and M.N.G. James for providing the structures before their release to the PDB. We also thank Peter Gund. The computations were done on Silicon Graphics Iris and NeXTstation workstations. The work at Harvard was partly supported by a grant from the National Science Foundation.

REFERENCES

1. Kleywegt, G. J., Bergfors, T., Senn, H., Le Motte, P., Gsell, B., Shudo, K., Jones, T. A. Crystal structure of cellular retinoic acid-binding proteins I and II in complex with all trans-retinoic acid and a synthetic retinoid. Structure 2:1241, 1994.
2. Šali, A., Overington, J. P., Johnson, M. S., Blundell, T. L. From comparisons of protein sequences and structures to protein modelling and design. TIBS 15:235–240, 1990.
3. Šali, A., Blundell, T. L. Definition of general topological equivalence in protein structures: A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. J. Mol. Biol. 212:403–428, 1990.
4. Šali, A., Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234:779–815, 1993.
5. Šali, A., Overington, J. P. Derivation of rules for comparative protein modeling from a database of protein structure alignments. Protein Sci. 3:1582–1596, 1994.
6. Šali, A., Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. In: "Protein Structure by Distance Analysis." Bohr, H., Brunak, S. (eds.). Amsterdam: IOS Press, 1994:64–86.
7. Šali, A. MODELLER-12, User's Guide and Programmer's Manual. Available by anonymous ftp from guitar.rockefeller.edu:pub/modeller/manuali.ps.z.
8. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F., Weng, J. Protein Data Bank. In: "Crystallographic Databases—Information, Content, Software Systems, Scientific Applications." Allen, F. H., Bergerhoff, G., Sievers, R. (eds.). Bonn/Cambridge/Chester: Data Commission of the International Union of Crystallography, 1987:107–132.
9. Šali, A., Matsumoto, R., McNeil, H. P., Karplus, M., Stevens, R. L. Three-dimensional models of four mouse mast cell chymases. Identification of proteoglycan-binding regions and protease-specific antigenic epitopes. J. Biol. Chem. 268:9023–9034, 1993.
10. Šali, A. Modeling mutations and homologous proteins. Curr. Opin. Biotech. 6:437–451, 1995.
11. Johnson, M. S., Srinivasan, N., Sowdhamini, R., Blundell, T. L. Knowledge-based protein modelling. CRC Crit. Rev. Biochem. Mol. Biol. 29:1–68, 1994.
12. Bajorath, J., Stenkamp, R., Aruffo, A. Knowledge-based model building of proteins: Concepts and examples. Protein Sci. 2:1798–1810, 1994.
13. Felsenstein, J. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791, 1985.
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. Sutcliffe, M. J., Haneef, I., Carney, D., Blundell, T. L. Knowledge based modelling of homologous proteins, Part I: Three dimensional frameworks derived from the simultaneous superposition of multiple structures. Prot. Eng. 1:377–384, 1987.
16. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., Karplus, M. CHARMM: A program

for macromolecular energy minimization and dynamics calculations. J. Comp. Chem. 4:187–217, 1983.
17. Laskowski, R. A., McArthur, M. W., Moss, D. S., Thornton, J. M. PROCHECK: A program to check the stereochemical quality of protein structures. J. Appl. Cryst. 26:283–291. 1993.
18. Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. Proteins 17:355–362, 1993.
19. Lüthy, R., Bowie, J. U., Eisenberg, D. Assessment of protein models with three-dimensional profiles. Nature (London) 356:83–85, 1992.
20. Summers, N. L., Karplus, M. Modeling of globular proteins: a distance-based search procedure for the construction of insertion/deletion regions and Pro → non-Pro mutations. J. Mol. Biol. 216:991–1016, 1990.
21. Ponder, J. W., Richards, F. M. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193:775–791, 1987.
22. Dunbrack, R. L., Karplus, M. Prediction of protein sidechain conformations from a backbone conformation dependent rotamer library. J. Mol. Biol. 230:543–571, 1993.
23. Overington, J., Johnson, M. S. Šali, A., Blundell, T. L. Tertiary structural constraints on protein evolutionary diversity; templates, key residues and structure prediction. Proc. R. Soc. London B 241:132–145, 1990.
24. Srinivasan, N., Blundell, T. L. An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure. Prot. Eng. 6:501–512, 1993.
25. Levitt, M. Molecular dynamics of native protein. II. Analysis and nature of motion. J. Mol. Biol. 168:621–657, 1983.
26. Kolinski, A., Skolnick, J. Monte Carlo simulations of protein folding. II. Application to protein A, ROP, and crambin. Proteins 18:353–366, 1994.
27. Fidelis, K., Stern, P. S., Bacon, D., Moult, J. Comparison of systematic search and database methods for constructing segments of protein structure. Protein Eng. 7:953–960, 1994.
28. Mattos, C., Petsko, G. A., Karplus, M. Analysis of two-residue turns in proteins. J. Mol. Biol. 238:733–747, 1994.
29. Borchert, T. V., Abagyan, R. A., Kishan, K. V. R., Zeelen, J. P., Wierenga, R. K. The crystal structure of an engineered monomeric triosephosphate isomerase, monoTim: The correct modelling of an eight residue loop. Structure 1:205–213, 1993.
30. Johnson, M. S., Overington, J. P. A structural basis for sequence comparisons: An evaluation of scoring methodologies. J. Mol. Biol. 233:716–738, 1993.
31. Jones, D. T., Taylor, W. R., Thornton, J. M. A new approach to protein fold recognition. Nature (London) 358: 86–89, 1992.
32. Godzik, A., Kolinski, A., Skolnick, J. Topology fingerprint approach to the inverse protein folding problem. J. Mol. Biol. 227:227–238, 1992.
33. Ohlendorf, D. H. Accuracy of refined protein structures. II. Comparison of four independently refined models of human interleukin 1β. Acta Cryst. D 50:808–812, 1994.
34. Clore, G. M., Robien, M.A., Gronenborn, A. M. Exploring the limits of precision and accuracy of protein structures determined by nuclear magnetic resonance spectroscopy. J. Mol. Biol. 231:82–102, 1993.
35. Zhao, D., Jardetzky, O. An assessment of the precision and accuracy of protein structures determined by NMR. J. Mol. Biol. 239:601–607, 1994.
36. Faber, H. R., Matthews, B. W. A mutant T4 lysozyme displays five different crystal conformations. Nature (London) 348:263–266, 1990.
37. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
38. Kraulis, P. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structure. J. Appl. Cryst. 24:946–950, 1991.
39. Matsumoto, R., Šali, A., Ghildyal, N., Karplus, M., Stevens, R.L. Packaging of proteases and proteoglycans in the granules of mast cells and other hematopoietic cells. A cluster of histidines in mouse mast cell protease-7 regulates its binding to heparin serglycin proteoglycan. J. Biol. Chem. 270:19524–19531, 1995.