HOMOLOGY PROTEIN STRUCTURE MODELING

Roberto Sánchez, Azat Ya. Badretdinov, Eric Feyfant and Andrej Šali

The Rockefeller University, 1230 York Avenue
New York, NY 10021, USA.

August 29, 1997

ABSTRACT

In the absence of a protein structure determined by X-ray crystallography or NMR spectroscopy, a 3D model of a given sequence can often be calculated by comparative modeling. In fact, this technique can be applied to an order of magnitude more protein sequences than the number of experimentally determined protein structures. A protein sequence with at least 40% identity to a known structure can sometimes be modeled automatically with an accuracy approaching that of a low resolution X-ray structure or a medium resolution NMR structure. This is so because the model closely resembles the template structure(s) and because the average structural differences between two proteins with more than 40% sequence identity are comparable to the differences between two independently determined low-resolution structures. Errors in comparative models include mistakes in the packing of side-chains, in conformation and shifts of the core segments and loops, and, most importantly, errors resulting from an incorrect alignment of the modeled sequence with related known structures. Despite the errors, the number of applications where comparative modeling has been proven useful is growing rapidly.

1. INTRODUCTION

Comparative or homology protein modeling uses experimentally determined protein structures (templates) to predict conformation of another protein with a similar amino acid sequence (target) [1–5]. This approach to modeling is possible because a small change in the protein sequence usually results in a small change in its three-dimensional (3D) structure [6, 7]. Comparative modeling is still the only modeling method that can provide models with an RMS error lower than 2Å (Section 3).

All current comparative modeling methods consist of four sequential steps [3, 5]. The first step is to identify the proteins with known 3D structures that are related to the target sequence. The second step is to align them with the target sequence and to pick those known structures that will be used as templates. The third step is to build the model for the target sequence given its alignment with the template structures. In the fourth step, the model is evaluated using a variety of criteria. If necessary, the alignment and model building are repeated until a satisfactory model is obtained.

The main difference between the different comparative modeling methods is in how the 3D model is calculated from a given alignment (step 3 above). The original and still the most widely used method is modeling by rigid body assembly [8–13]. The method constructs the model from a few rigid bodies that include core regions, loops and side-chains, all of which are obtained from dissecting related structures. The assembly of the model involves calculating a framework, which is defined as the average of the template C_α atoms in the conserved regions of the fold, and then fitting the rigid bodies on the framework. Another family of methods, modeling by segment matching, relies on approximate positions of conserved atoms from the templates to calculate the coordinates of other atoms [14–17]. This is achieved by the use of a database of short segments of protein structure, energy or geometry rules, or some combination of these criteria. The third group of methods, modeling by satisfaction of spatial restraints, uses either distance geometry [18–20] or optimization techniques [21, 22] to satisfy spatial restraints obtained from the alignment of the target sequence with similar templates of known structure. Some available software packages for comparative modeling are listed in Table 1. In addition to the methods for modeling the whole fold, numerous other techniques for predicting loops [23] and side-chains [24] on a given backbone have also been described. These methods can often be used in combination with each other and with comparative modeling techniques.

Comparative modeling, even if less accurate than experimental methods, can be helpful in proposing and testing hypotheses in molecular biology, such as hypotheses about the location of ligand binding sites [25], substrate specificity [26], and drug design [27]. It can also provide starting models in X-ray crystallography [28] and NMR spectroscopy [29]. An exhaustive survey of many comparative modeling studies is given in [1]. However, the main challenge to comparative protein modeling is posed by the genome sequencing projects. In a few years, the genome projects will have provided us with the amino acid sequences of more than 500,000 proteins — the catalysts, inhibitors, messengers, receptors, transporters, and building blocks of the living organisms. The full potential of the genome projects will only be realized once we assign and understand the function of these new proteins. The understanding, modification and manipulation of protein function generally require knowledge of the 3D structure of a protein at the atomic level. Unfortunately, experimental methods for protein structure determination are time consuming and not successful with all proteins; consequently, 3D structures have been determined for only a tiny fraction of proteins for which the amino acid sequence is known. However, for many protein sequences, comparative modeling can provide a useful 3D model. In fact, about one third of the 239,207 known protein sequences (GENPEPT, 23 June, 1997) are related to at least one of the 5,698 known protein structures (Brookhaven Protein Databank [30], 30 July, 1997) [31]. Thus, the number of sequences that can be modeled relatively accurately at this moment is an order of magnitude larger than the number of experimentally determined protein structures. Furthermore, the usefulness of comparative modeling is steadily increasing because genome projects are producing more sequences and because novel protein folds are being determined experimentally. It has been estimated that there are approximately 1,000 different protein fold families, one third of which have already been structurally defined [31–33]. Assuming the current growth rate in the number of known protein structures, the structure of at least one member of most protein fold families will be determined in only about 6 years (Figure 4C in [31]), thus allowing comparative modeling to be applicable to most of the protein sequences (domains) at that time.

This review describes our own approach to comparative protein structure modeling and its evaluation. ←Tab.1

2. COMPARATIVE PROTEIN STRUCTURE MODELING BY SATISFACTION OF SPATIAL RESTRAINTS

We developed an automated approach to comparative protein modeling that is based on satisfaction of spatial restraints [21, 34–38] (Figure 1). It is implemented in the computer program MODELLER which is freely available to academic researchers *via* World Wide Web at URL `http://guitar.rockefeller.edu`. Graphical interfaces to MODELLER are provided by QUANTA, INSIGHTII, and GENEEXPLORER (MSI, San Diego, CA; e-mail `blp@msi.com`).

The comparative modeling by MODELLER begins with an alignment of the target sequence with related known 3D structures. The output, obtained without any user intervention, is a 3D model for the target sequence containing all mainchain and sidechain non-hydrogen atoms. In the first phase of the modeling process, many distance and dihedral angle restraints on the target sequence are derived from its alignment with template 3D structures (Figure 2). The form of these restraints was obtained from the statistical analysis of the relationships between similar structures. This analysis relied on a database of 105 family alignments that include 416 proteins with known 3D structure [36]. By scanning the database, tables quantifying various correlations were obtained, such as the correlations between two equivalent $C_\alpha - C_\alpha$ distances, or between equivalent mainchain dihedral angles from two related proteins [21]. These relationships were expressed as conditional probability density functions (pdf's) and can be used directly as spatial restraints. For example, probabilities for different values of the mainchain dihedral angles are calculated from the type of a residue considered, from mainchain conformation of an equivalent residue, and from sequence similarity between the two proteins. Another example is the pdf for a certain C_α–C_α distance given equivalent distances in two related protein structures (Figure 2). An important feature of the method is that the forms of spatial restraints are obtained empirically, from a database of protein structure alignments. Next, the spatial restraints and CHARMM energy terms enforcing proper stereochemistry [39] are combined into an objective function. Finally, the model is obtained by optimizing the objective function in Cartesian space. The optimization is carried out by the use of the variable target function method [40] employing methods of conjugate gradients and molecular dynamics with simulated annealing [41] (Figure 3). Several slightly different models can be calculated by varying the initial structure and the variability among these models can be used to estimate the errors in the corresponding regions of the fold.

←Fig.1
←Fig.2
←Fig.3

Because the modeling by satisfaction of spatial restraints can use many different types of information about the target sequence, it is perhaps the most promising of all comparative modeling techniques. For example, restraints could be provided by rules for secondary structure packing [42], analyses of hydrophobicity [20] and correlated mutations [43], empirical potentials of mean force [44], NMR experiments [29], cross-linking experiments [45], image reconstruction in electron microscopy [46], site-directed mutagenesis [47], fluorescence spectroscopy, intuition, *etc.* In this way, a comparative model, especially in the difficult cases, could be improved by making it consistent with available experimental data and/or with more general knowledge about protein structure.

Comparative modeling by satisfaction of spatial restraints has been used by us and our collaborators to study proteoglycan binding, antigenicity, and substrate specificity of mouse mast cell chymases and tryptases [25, 48–51], the phospholipid binding site of β_2-glycoprotein I [52], substrate specificity of brain lipid binding protein [26], the overall electrostatic properties and DNA binding of *E. coli* transcript cleavage factors [53], proteins related to the δ' subunit of the clamp-loader complex of *E. coli* DNA polymerase III [54], and the sequence conservation patterns in the phosphoglucomutase-like gene of *Staphylococcus aureus* [55].

3. ERRORS IN HOMOLOGY-DERIVED PROTEIN STRUCTURE MODELS

Recently, protein modelers were challenged again to model sequences with unknown 3D structure and to submit their models to the second "Meeting on Critical Assessment of Techniques for Protein Structure Prediction" (CASP) in Asilomar in December of 1996 (URL `http://iris4.carb.nist.gov/casp2/`). At the same time, the 3D structures of the prediction targets were being determined by X-ray crystallography or NMR methods. They only became available after the models were calculated. Thus, it was possible to test the modeling methods objectively [38, 56].

It turned out that the best comparative techniques can generally produce models with good stereochemistry and overall structural accuracy that is slightly higher than the similarity between the template and the actual target structures, when the modeling alignment is correct. There appeared to be two modest improvements relative to the results at the first CASP meeting in 1994 [57]: They were a consequence of better alignments resulting from more careful manual editing and of better techniques for modeling of the insertions shorter than about 9 residues.

The errors in comparative models can be divided into four categories [38, 56]: (1) Errors in sidechain packing (Figure 4). (2) Distortions or shifts of a region that is aligned correctly with the templates (Figure 5). (3) Distortions or shifts of a region that does not have an equivalent segment in any of the templates (Figure 6). (4) Distortions or shifts of a region that is aligned incorrectly with the templates (Figure 7).

←Fig.4
←Fig.5
←Fig.6
←Fig.7

Errors 2–4 are relatively infrequent when sequences with more than 40% identity to the templates are modeled. For example, in such a case, approximately 90% of the mainchain atoms are likely to be modeled with an RMS error of about 1Å. In this range of sequence similarity, the alignment is mostly straightforward to construct, there are not many gaps, and structural differences between the proteins are usually limited to loops and side-chains. When sequence identity is between 30 and 40%, the structural differences become larger, and the gaps in the alignment are more frequent and longer. As a result, the main-chain RMS error rises to about 1.5Å for about 80% of residues. The rest of the residues are modeled with large errors because the methods generally cannot model structural distortions and rigid body shifts, and cannot recover from misalignments. Below 40% sequence identity, misalignments and insertions in the target sequence become the major problems. Insertions longer than about 9 residues cannot be modeled accurately at this time, while shorter loops frequently can be modeled successfully [23, 58–62]. When sequence identity drops below 30%, the main problem becomes the identification of related templates and their alignment with the sequence to be modeled. In general, it can be expected that about 20% of residues will be misaligned, and consequently incorrectly modeled with an error larger than 3Å, at this level of sequence similarity [63]. This is a serious impediment for comparative modeling because it appears that at least one half of all related protein pairs are related at less than 40% sequence identity [64].

When there are alignment errors in the template – target alignment used for modeling and when the correct, structure-based template – target alignment is used for comparing the template with the actual target structure, the target structure is frequently more similar to the closest template structure than to the model. In contrast, if the modeling target – template alignment is used in evaluating the similarity between the actual target structure and the template, the target structure is generally closer to the model than to the template [56]. As a result, using a model is generally better than using the template structure even when the alignment is incorrect because the actual target structure, and therefore the correct template – target alignment, are not available in practical modeling applications.

To put the errors into perspective, we list the differences among experimentally determined structures of the same protein. The 1Å accuracy of main-chain atom positions corresponds to X-ray structures defined at a low-resolution of about 2.5Å and with an R-factor of about 25% [65], as well as to medium-resolution NMR structures determined from 10 inter-proton distance restraints per residue [66, 67]. Similarly, differences between the highly refined X-ray and NMR structures of the same protein also tend to be about 1Å[66]. Changes in the environment (*e.g.*, oligomeric state, crystal packing, solvent, ligands) can also have a significant effect on the structure [68]. Overall, comparative modeling based on templates with more than 40% identity is almost as good, 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.

4. DISCUSSION

Future improvements of comparative modeling should aim to model proteins with lower similarities to known structures (*e.g.*, less than 30% sequence identity), to increase the accuracy of the models, and to make modeling fully automated. The improvements are likely to include simultaneous optimization of side-chain and backbone conformations in side-chain modeling, simultaneous optimization of a loop and its environment in loop modeling, and simultaneous optimization of the alignment and the model. At the same time, better potential functions and possibly better optimizers 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 homology-derived scoring function, such as that of MODELLER [21], could be one way of achieving this goal [69–71]. This is a difficult problem, as illustrated by the fact that no present force field or potential of mean force can produce a model with a main-chain RMSD from the X-ray structure smaller than about 1Å, even when the starting conformation is the X-ray structure itself. For example, molecular dynamics simulations in solvent generally have a main-chain RMSD of more than 1Å and the most detailed lattice folding simulations result in models with an RMS error larger than 2Å[72]. Since most of the main-chain atoms in two homologs with at least 40% sequence identity usually superpose with an RMSD of about 1Å, 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.

The major factor that limits the use of comparative modeling in the cases of less than 30% sequence identity is the alignment problem. In principle, the alignment can be derived by any of the sequence or sequence/structure alignment methods, but in practice even careful manual editing frequently results in significant alignment errors. At 30% sequence identity, the fraction of incorrectly aligned residues is about 20% and this number rises sharply with further decrease in sequence similarity [63]; an additional complication is that even structure/structure comparison may not result in a unique alignment for proteins with less than about 25% identity [73, 74]. This limits the usefulness of comparative modeling because no current modeling technique can recover from an incorrect input alignment. It would appear that profile matching [75] and threading methods [76–78] are a natural solution to the alignment problem in comparative modeling. However, while these techniques are successful in identifying related folds, they appear to be somewhat less successful in generating correct alignments. To reduce the errors in the model stemming from the alignment errors, iterative changes in the alignment during the calculation of the model are needed [54, 56].

Even though comparative modeling needs significant improvements, it is already a mature technique that can be used 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 an increasingly important tool for biologists who seek to understand and control normal and disease related processes in living organisms.

ACKNOWLEDGMENTS

AŠ is a Sinsheimer Scholar. RS is a Howard Hughes Medical Institute predoctoral fellow. EF is a fellow of Association pour le Developpement de l'Enseignement et des Recherches pour la Region Provence-Alpes-Cote d'Azur. Grants GM 54762 from NIH and BIR-9601845 from NSF are also acknowledged. This review is based on [3, 5, 38, 56, 79].

References

[1] Johnson, M. S., Srinivasan, N., Sowdhamini, R., and Blundell, T. L., CRC Crit. Rev. Biochem. Mol. Biol. **29** (1994) 1.

[2] Bajorath, J., Stenkamp, R., and Aruffo, A., Protein Sci. **2** (1994) 1798.

[3] Šali, A., Curr. Opin. Biotech. **6** (1995) 437.

[4] Rost, B. and Sander, C., Annu. Rev. Biophys. Biomol. Struct. **25** (1996) 113.

[5] Sánchez, R. and Šali, A., Curr. Opin. Str. Biol. **7** (1997) 206.

[6] Lesk, A. M. and Chothia, C. H., Philos. Trans. R. Soc. London Ser. B **317** (1986) 345.

[7] Hubbard, T. J. P. and Blundell, T. L., Protein Eng. **1** (1987) 159.

[8] Browne, W. J. et al., J. Mol. Biol. **42** (1969) 65.

[9] Greer, J., J. Mol. Biol. **153** (1981) 1027.

[10] Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E., and Thornton, J. M., Nature **326** (1987) 347.

[11] Vriend, G., J. Mol. Graph. **8** (1990) 52.

[12] Peitsch, M. C. and Jongeneel, C. V., Int. Immunol. **5** (1993) 233.

[13] Bruccoleri, R. E., Molecular Simulation **10** (1993) 151.

[14] Jones, T. H. and Thirup, S., EMBO J. **5** (1986) 819.

[15] Unger, R., Harel, D., Wherland, S., and Sussman, J. L., Proteins **5** (1989) 355.

[16] Claessens, M., Cutsem, E. V., Lasters, I., and Wodak, S., Protein Eng. **4** (1989) 335.

[17] Levitt, M., J. Mol. Biol. **226** (1992) 507.

[18] Havel, T. F. and Snow, M. E., J. Mol. Biol. **217** (1991) 1.

[19] Srinivasan, S., March, C. J., and Sudarsanam, S., Protein Sci. **2** (1993) 227.

[20] Aszódi, A. and Taylor, W. R., Folding and Design **1** (1996) 325.

[21] Šali, A. and Blundell, T. L., J. Mol. Biol. **234** (1993) 779.

[22] Brocklehurst, S. M. and Perham, R. N., Protein Sci. **2** (1993) 626.

[23] van Vlijmen, H. W. T. and Karplus, M., J. Mol. Biol. **267** (1997) 975.

[24] Bower, M. J., Cohen, F. E., and Dunbrack, R. L., J. Mol. Biol. **267** (1997) 1268.

[25] Matsumoto, R., Šali, A., Ghildyal, N., Karplus, M., and Stevens, R. L., J. Biol. Chem. **270** (1995) 19524.

[26] Xu, L. Z., Sánchez, R., Šali, A., and Heintz, N., J.Biol.Chem. **271** (1996) 24711.

[27] Ring, C. S. et al., Proc. Natl. Acad. Sci. USA **90** (1993) 3583.

[28] Carson, M., Bugg, C. E., Delucas, L., and Narayana, S., Acta Crystallogr. **D50** (1994) 889.

[29] Sutcliffe, M. J., Dobson, C. M., and Oswald, R. E., Biochemistry **31** (1992) 2962.

[30] Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T., and Weng, J., Protein data bank, in *Crystallographic databases — Information, content, software systems, scientific applications*, edited by Allen, F. H., Bergerhoff, G., and Sievers, R., pages 107–132, Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, 1987.

[31] Holm, L. and Sander, C., Science **273** (1996) 595.

[32] Chothia, C., Nature **360** (1992) 543.

[33] Orengo, C. A., Jones, D. T., and Thornton, J. M., Nature **372** (1994) 631.

[34] Šali, A., Overington, J. P., Johnson, M. S., and Blundell, T. L., TIBS **15** (1990) 235.

[35] Šali, A. and Blundell, T. L., J. Mol. Biol. **212** (1990) 403.

[36] Šali, A. and Overington, J., Protein Sci. **3** (1994) 1582.

[37] Šali, A. and Blundell, T. L., Comparative protein modelling by satisfaction of spatial restraints, in *Protein Structure by Distance Analysis*, edited by Bohr, H. and Brunak, S., pages 64–86, IOS Press, Amsterdam, 1994.

[38] Šali, A., Potterton, L., Yuan, F., van Vlijmen, H., and Karplus, M., Proteins **23** (1995) 318.

[39] Brooks, B. R. et al., J. Comp. Chem. **4** (1983) 187.

[40] Braun, W. and Gõ, N., J. Mol. Biol. **186** (1985) 611.

[41] Clore, G. M., Brünger, A. T., Karplus, M., and Gronenborn, A. M., J. Mol. Biol. **191** (1986) 523.

[42] Taylor, W. R., Protein Eng. **6** (1993) 593.

[43] Gobel, U., C., S., Schneider, R., and Valencia, A., Proteins **18** (1994) 309.

[44] Sippl, M. J., J. Mol. Biol. **213** (1990) 859.

[45] Rossi, V. et al., Biochemistry **34** (1995) 7311.

[46] Neil, K. J., Nucleic Acids Research **24** (1995) 1472.

[47] Boissel, J. P., Lee, W. R., Presnell, S. R., Cohen, F. E., and Bunn, H. F., J. Biol. Chem. **268** (1993) 15983.

[48] Šali, A., Matsumoto, R., McNeil, H. P., Karplus, M., and Stevens, R. L., J. Biol. Chem. **268** (1993) 9023.

[49] Ghildyal, N. et al., J. Exp. Med. **184** (1996) 1061.

[50] Hunt, J. E. et al., J. Biol. Chem., in press (1997).

[51] Huang, C. et al., submitted (1997).

[52] Sheng, Y., Šali, A., Herzog, H., Lahnstein, J., and Krilis, S., J. Immunol. **157** (1996) 3744.

[53] Koulich, D. et al., J. Biol. Chem. **272** (1997) 7201.

[54] Guenther, B., Onrust, R., Šali, A., O'Donnell, M., and Kuriyan, J., Cell, in press (1997).

[55] Wu, S., de Lencastre, H., Šali, A., and Tomasz, A., Microbial Drug Resistance **2** (1996) 277.

[56] Sánchez, R. and Šali, A., Proteins, in press (1997).

[57] Mosimann, S., Meleshko, R., and James, M. N. G., Proteins **23** (1995) 301.

[58] Mattos, C., Petsko, G. A., and Karplus, M., J. Mol. Biol. **238** (1994) 733.

[59] Fidelis, K., Stern, P. S., Bacon, D., and Moult, J., Protein Eng. **7** (1994) 953.

[60] Borchert, T. V., Abagyan, R. A., Kishan, K. V. R., Zeelen, J. P., and Wierenga, R. K., Structure **1** (1993) 205.

[61] Collura, V., Higo, J., and Garnier, J., Protein Sci. **2** (1993) 1502.

[62] Bassolino-Klimas, D., Bruccoleri, R. E., and Subramaniam, S., Protein Sci. **1** (1992) 1465.

[63] Johnson, M. S. and Overington, J. P., J. Mol. Biol. **233** (1993) 716.

[64] Rost, B., Folding & Design **2** (1997) S19.

[65] Ohlendorf, D. H., Acta Cryst. **D50** (1994) 808.

[66] Clore, G. M., Robien, M. A., and Gronenborn, A. M., J. Mol. Biol. **231** (1993) 82.

[67] Zhao, D. and Jardetzky, O., J. Mol. Biol. **239** (1994) 601.

[68] Faber, H. R. and Matthews, B. W., Nature **348** (1990) 263.

[69] Sippl, M. J., Ortner, M., Jaritz, M., Lackner, P., and Flöckner, H., Folding & Design **1** (1996) 275.

[70] DeBolt, S. E. and Skolnick, J., Protein Eng. **9** (1996) 937.

[71] Melo, F. and Feytmans, E., J. Mol. Biol. **267** (1997) 207.

[72] Kolinski, A. and Skolnick, J., Proteins **18** (1994) 353.

[73] Zu-Kang, F. and Sippl, M. J., Folding and Design **1** (1996) 123.

[74] Godzik, A., Protein Science **5** (1996) 1325.

[75] Bowie, J. U., Lüthy, R., and Eisenberg, D., Science **253** (1991) 164.

[76] Finkelstein, A. V. and Reva, B. A., Nature **351** (1991) 497.

[77] Jones, D. T., Taylor, W. R., and Thornton, J. M., Nature **358** (1992) 86.

[78] Godzik, A., Kolinski, A., and Skolnick, J., J. Mol. Biol. **227** (1992) 227.

[79] Sánchez, R. and Šali, A., Journal of Molecular Structure (Theochem) **in press** (1997).

[80] Havel, T. F., Mol. Simulation **10** (1993) 175.

[81] Sutcliffe, M. J., Haneef, I., Carney, D., and Blundell, T. L., Protein Eng. **1** (1987) 377.

[82] Kleywegt, G. J. et al., Structure **2** (1994) 1241.

[83] Kabsch, W. and Sander, C., Biopolymers **22** (1983) 2577.

Table 1

Some available software packages for comparative modeling of whole proteins. Method key (see also Introduction): 1, comparative modeling by assembly of rigid bodies; 2, comparative modeling by segment matching; 3, comparative modeling by satisfaction of spatial restraints. SYBYL includes COMPOSER; QUANTA and InsightII include MODELLER as well as in-house algorithms for modeling by rigid body assembly; InsightII also includes CONSENSUS [80]. SWISS-MOD is an Internet server for comparative modeling that takes either the target sequence or its alignment with known structures as input. There are many additional programs that specialize in modeling of sidechains or loops only.

Program	*Availability*	*Address*	*Method*	*Reference*
COMPOSER	public	`http://felix.bioc.cam.ac.uk/soft-base.html`	1	[81]
CONGEN	public	`email: bruc@dino.squibb.com`	1	[13]
DRAGON	public	`http://www.nimr.mrc.ac.uk/~mathbio/a-aszodi/dragon.html`	3	[20]
MODELLER	public	`http://guitar.rockefeller.edu/modeller/modeller.html`	3	[21]
NAOMI	public	`http://www.ocms.ox.ac.uk./~smb/Software/N_details/naomi.html`	3	[22]
WHAT IF	public	`http://swift.embl-heidelberg.de/whatif/`	1	[11]
InsightII	commercial	`http://www.msi.com/`	1, 3	MSI, San Diego
LOOK	commercial	`http://www.mag.com/`	2	[17]
QUANTA	commercial	`http://www.msi.com/`	1, 3	MSI, San Diego
SYBYL	commercial	`http://www.tripos.com/`	1, 3	Tripos, St. Louis
SWISS-MOD	public server	`http://www-isrec.unil.ch/SWISS-MODEL.html`	1	[12]

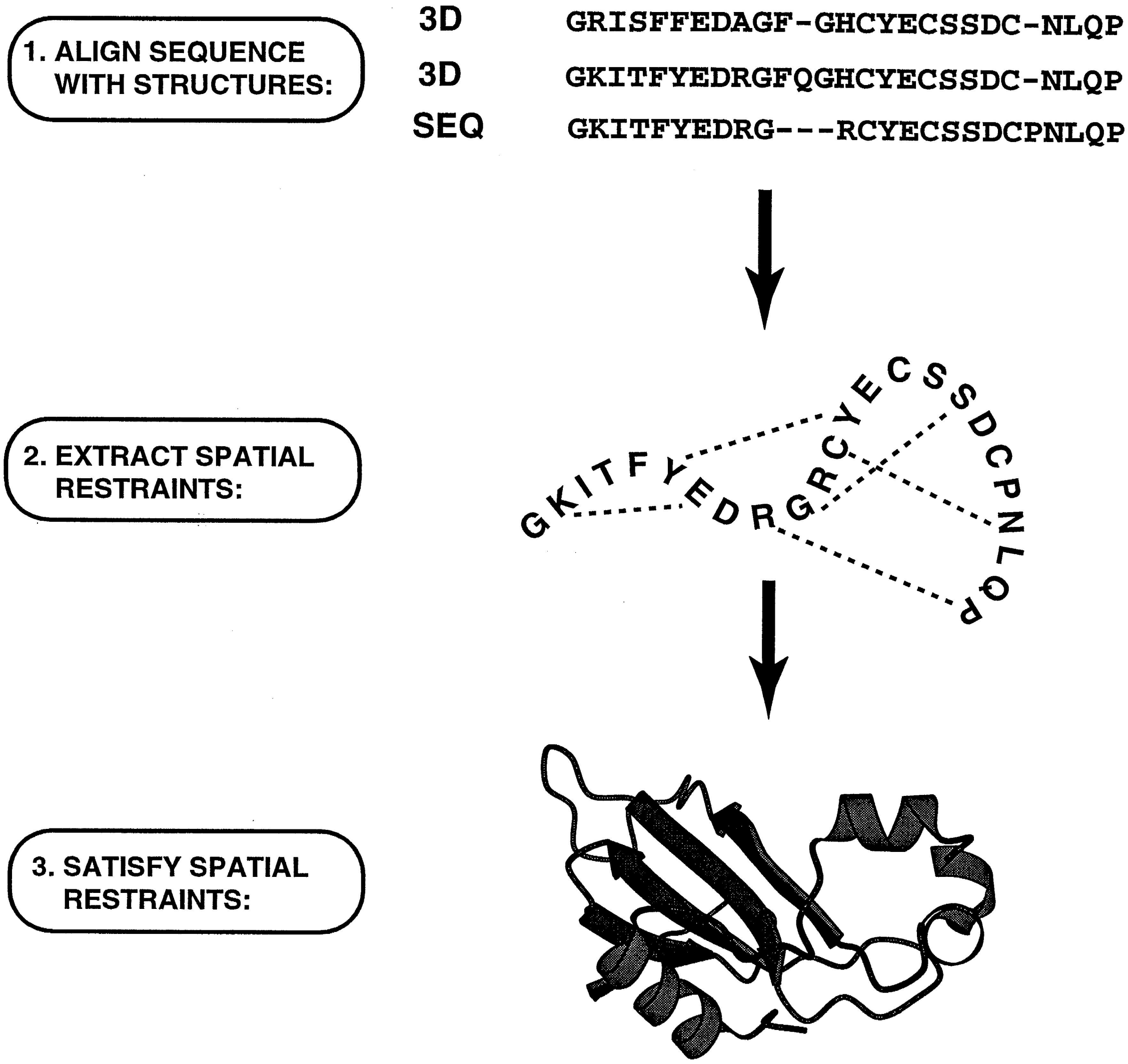
1. ALIGN SEQUENCE WITH STRUCTURES:
3D GRISFFEDAGF-GHCYECSSDC-NLQP
3D GKITFYEDRGFQGHCYECSSDC-NLQP
SEQ GKITFYEDRG---RCYECSSDCPNLQP
2. EXTRACT SPATIAL RESTRAINTS:
GKITFYEDRGRCYECSSDCPNLQP
3. SATISFY SPATIAL RESTRAINTS:

Figure 1: Comparative protein modeling by satisfaction of spatial restraints. First, the sequence to be modeled (target) is aligned with the known 3D structures (templates). Second, a large number of restraints (dashed lines) on distances and dihedral angles in the target sequence are extracted from the alignment. Third, the 3D model is obtained by satisfying all the restraints as well as possible. Reprinted with permission from [79].

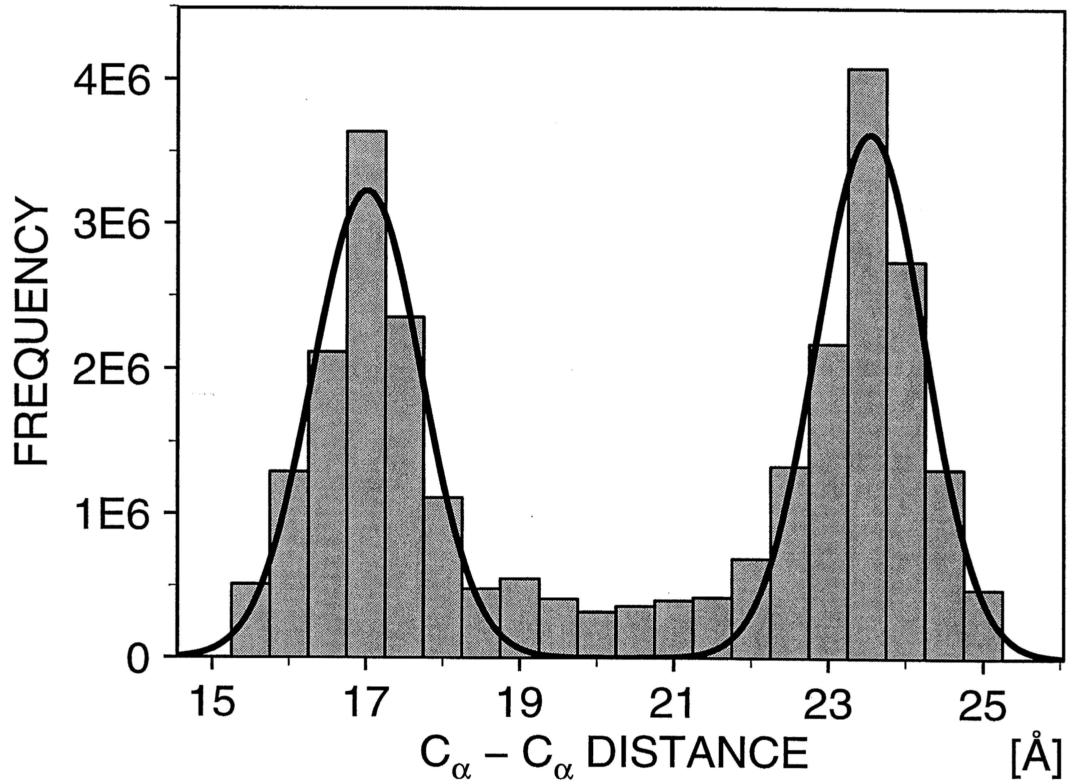

Figure 2: Sample spatial restraint. A restraint on a given C_α–C_α distance, d, is expressed as a conditional probability density function that depends on two other equivalent distances ($d' = 17.0$ and $d'' = 23.5$): $p(d/d', d'')$. The restraint (continuous line) is obtained by least-squares fitting a sum of two Gaussian functions to the histogram, which in turn is derived from the database of alignments of protein structures. In practice, more complicated restraints are used that depend on additional information, such as similarity between the proteins, solvent accessibility, and distance from a gap in the alignment. Reprinted with permission from [79].

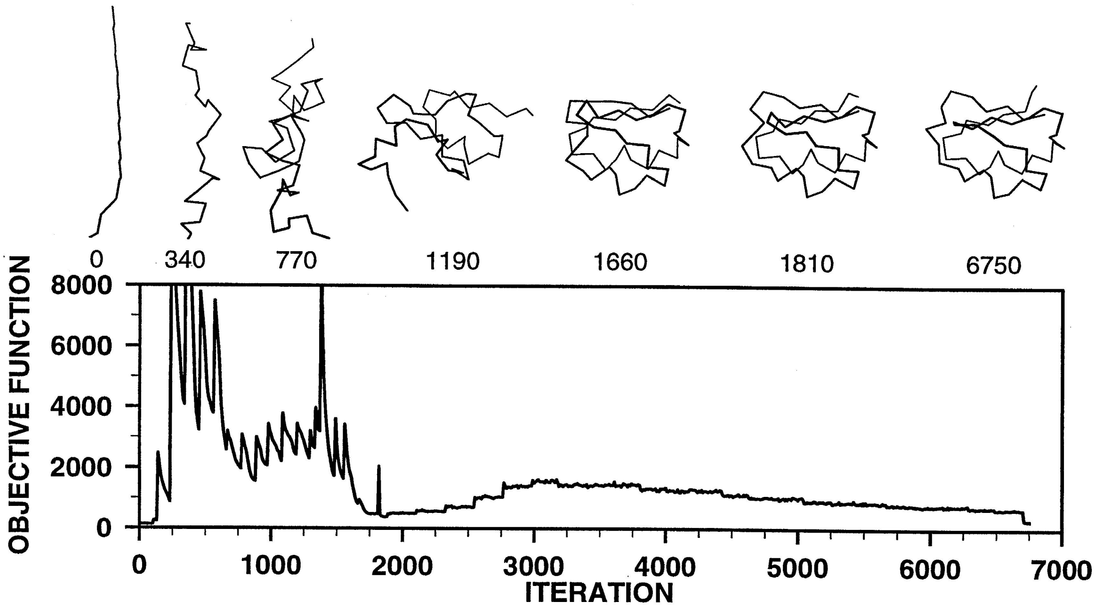

Figure 3: Optimization of the objective function. Optimization of the objective function (curve) starts with a random or distorted template structure. The iteration number is indicated below each sample structure. The first ~ 2,000 iterations correspond to the variable target function method [40] relying on the conjugate gradients technique. This approach first satisfies sequentially local restraints and slowly introduces longer range restraints until the complete objective function is optimized. In the last 4,750 iterations, molecular dynamics with simulated annealing is used to refine the model [41]. CPU time needed to generate one model is about 5 min for a 250 residue protein on a Silicon Graphics Inc. R10000 workstation. Reprinted with permission from [79].

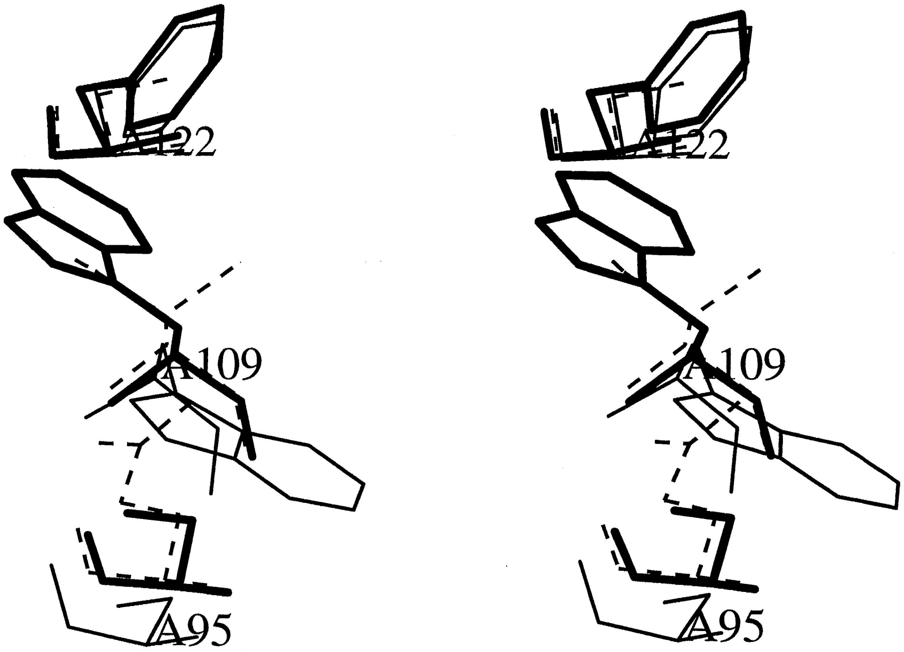

Figure 4: Errors in sidechain packing. The Trp 109 residue in the crystallographically determined structure of mouse cellular retinoic acid binding protein I [82] (thin line) is compared with its model (thick line), and with the template mouse adipocyte lipid-binding protein (the Brookhaven code 1LIF) (dashed line). Reprinted with permission from [79].

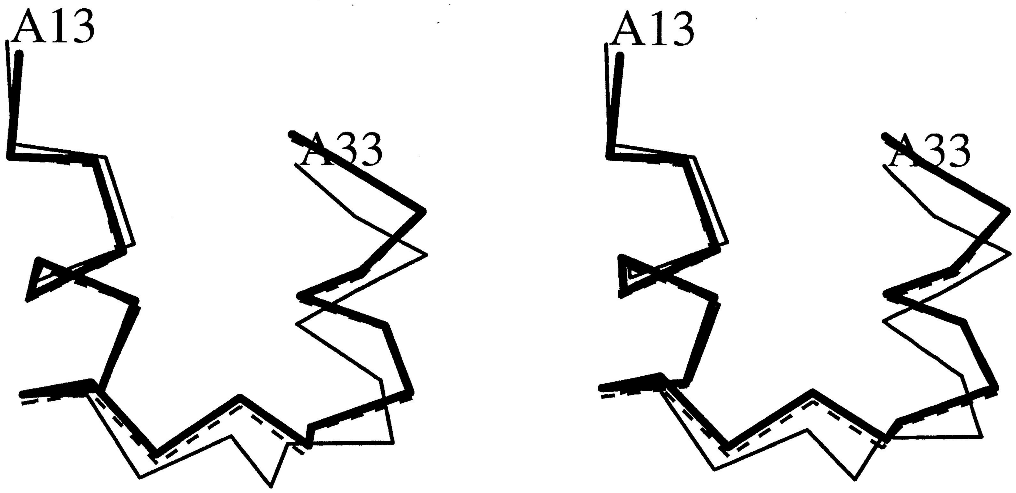

Figure 5: Distortions and shifts in correctly aligned regions. A region in the crystallographically determined structure of mouse cellular retinoic acid binding protein I [82] (thin line) is compared with its model (thick line), and with the template fatty acid binding protein (the Brookhaven code 2HMB) (dashed line). Reprinted with permission from [79].

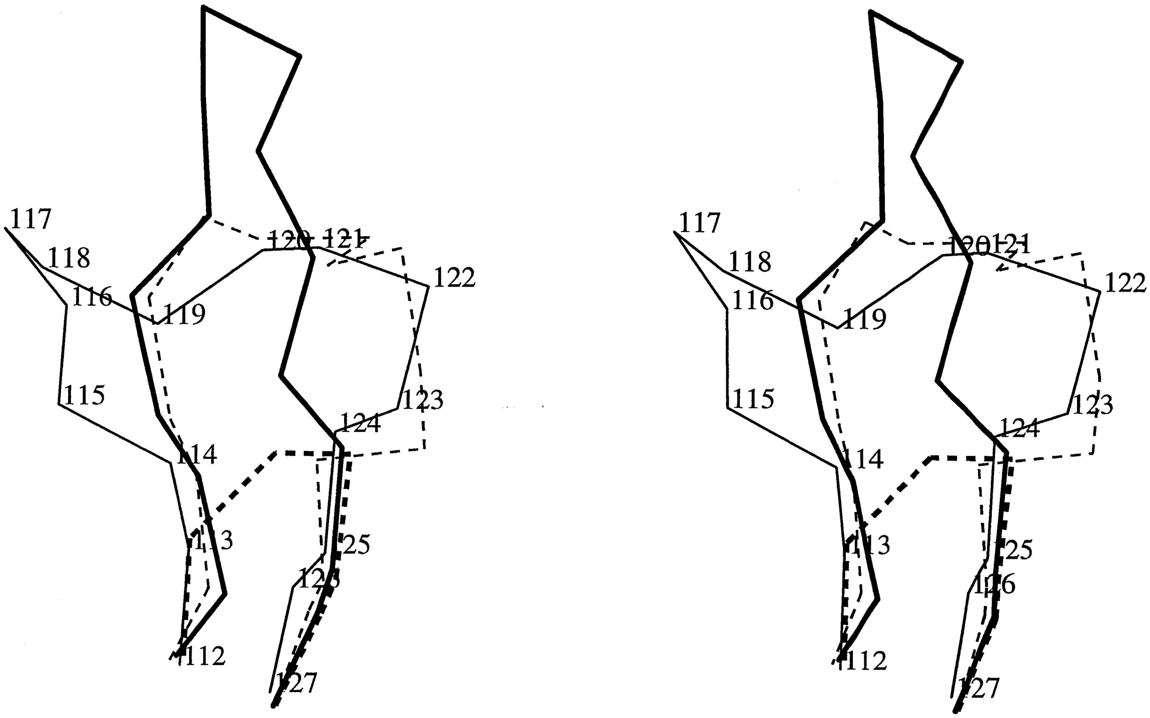

Figure 6: Errors in unaligned regions. Stereo plot of the C_α trace of the 112–127 loop is shown for the X-ray structure of human eosinophil neurotoxin (S.C. Mosimann, D. Newton, R. Youle, and M.N.G. James, in preparation) (continuous thin line), its model (thick line), and the template ribonuclease A structure (residues 111–117; thick dashed line). Reprinted with permission from [38].

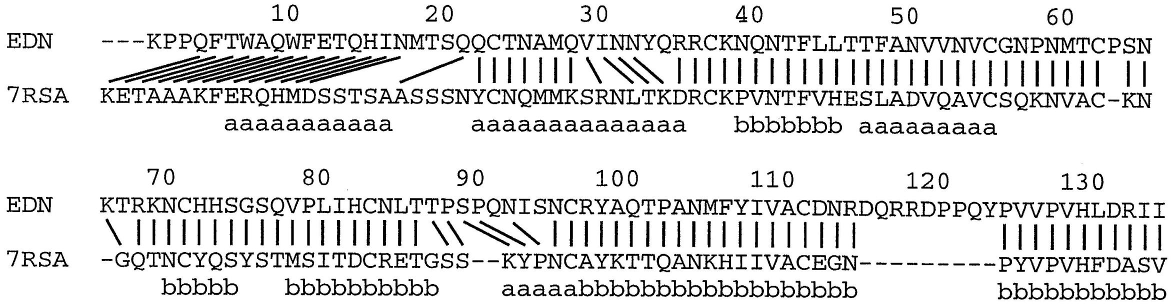
10 20 30 40 50 60
EDN ---KPPQFTWAQWFETQHINMTSQQCTNAMQVINNYQRRCKNQNTFLLTTFANVVNVCGNPNMTCPSN
7RSA KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHESLADVQAVCSQKNVAC-KN
aaaaaaaaaaa aaaaaaaaaaaaaa bbbbbbb aaaaaaaaa
70 80 90 100 110 120 130
EDN KTRKNCHHSGSQVPLIHCNLTTPSPQNISNCRYAQTPANMFYIVACDNRDQRRDPPQYPVVPVHLDRII
7RSA -GQTNCYQSYSTMSITDCRETGSS--KYPNCAYKTTQANKHIIVACEGN---------PYVPVHFDASV
bbbbb bbbbbbbbbb aaaaabbbbbbbbbbbbbbbbb bbbbbbbbbbb

Figure 7: Errors in the sequence alignment of human eosinophil neurotoxin and ribonuclease A. Automatically derived sequence alignment is shown. The black lines show correct equivalences, that is residues whose C_α atoms are within 5Å of each other in the optimal least-squares superposition of the two X-ray structures. The bottom line indicates helices (a) and strands (b), as assigned in the human eosinophil neurotoxin structure by program DSSP [83]. Reprinted with permission from [38].