ELSEVIER

Journal of Molecular Structure (Theochem) 398–399 (1997) 489–496

THEO CHEM

Comparative protein structure modeling as an optimization problem[1]

Roberto Sánchez, Andrej Šali*

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

Abstract

Comparative or homology protein modeling uses experimentally determined protein structures to predict the conformation of another protein with a similar amino acid sequence. This technique can produce useful models for about an order of magnitude more protein sequences than there have been structures determined by experiment. Here, we review our approach to comparative protein modeling. In this approach, the three-dimensional model is calculated by satisfying spatial restraints extracted from an alignment of the sequence to be modeled with related known structures. We examine the types of errors in the resulting models and discuss some of the potential advantages of formulating comparative modeling as an optimization problem. © 1997 Elsevier Science B.V.

Keywords: Comparative protein modeling; Optimization; Molecular conformation; Protein structure

1. Introduction

In a few years, the genome projects will have provided us with amino acid sequences of approximately 500 000 proteins. The full potential of the genome projects will only be realized once we can assign, understand, and manipulate the function of these new proteins. Such control of protein function generally requires knowledge of protein three-dimensional (3D) structure. Unfortunately, experimental methods for protein structure determination, such as X-ray crystallography and NMR spectroscopy, 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. In the absence of a high-resolution protein structure determined by X-ray crystallography or NMR spectroscopy, a useful 3D model of a given sequence can often be calculated by comparative modeling.

Comparative or homology protein modeling uses experimentally determined protein structures (templates) to predict the conformation of another protein with a similar amino acid sequence (target) [1–6]. This is possible because a small change in the sequence usually results in a small change in the 3D structure [7]. All comparative modeling methods begin with an alignment between the target and templates. The main difference between the different comparative modeling methods is in how the 3D model is calculated from a given alignment. The oldest and still the most widely used method is modeling by rigid body assembly [3]. The method constructs the model from a few core regions, loops and side-chains, which are obtained from dissected related structures. This assembly involves fitting the rigid bodies on the framework, which is defined as the average of the C_α

* Corresponding author.

[1] Presented at WATOC '96, Jerusalem, Israel, 7–12 July 1996.

PII S0166-1280(96)04971-8

atoms in the conserved regions of the fold. 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 [8]. 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, satisfies spatial restraints obtained from the alignment of the target sequence with homologous templates of known structure [9,10]. As this restraint-based modeling can use many different types of information about the target sequence, it is perhaps the most promising of all comparative modeling techniques.

A recent version of the Protein Information Resource database of protein sequences (PIR 45) contained 134 896 entries on 30 August 1995 [11]. In contrast, the Brookhaven Protein Databank of experimentally determined protein structures contained only 3836 entries on 22 September 1995 [12]. Since about one-third of known sequences appear to be related to at least one known structure [13], the number of sequences that can be modeled 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. Thus, comparative modeling will be an increasingly important tool for biologists who seek to understand and control normal and disease-related processes in living organisms. Typical applications include the formulation and testing of hypotheses about ligand binding sites [14,15], substrate specificity [16], and drug design [17]; it can also provide starting models in X-ray crystallography [18] and NMR spectroscopy [19].

In the subsequent sections, we first outline our approach to comparative protein structure modeling, which is implemented in the computer program MODELLER (Section 2). We then summarize the errors in the models derived by the current version of MODELLER, version 3 (Section 3). Finally, we outline the tools for defining and optimizing objective functions in MODELLER (Section 4) and discuss the potential advantages of formulating comparative modeling as an optimization problem (Section 5).

2. Comparative protein modeling by satisfaction of spatial restraints

We developed an automated approach to comparative protein modeling that is based on satisfaction of spatial restraints [9,20–24] (Fig. 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 and INSIGHTII (MSI, San Diego, CA; e-mail: blp@biosym.com).

The comparative modeling procedure 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 main-chain and side-chain 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 (Fig. 2). The form of these restraints was obtained from the statistical analysis of the relationships between homologous structures. This analysis relied on a database of 105 family alignments that include 416 proteins with known 3D structure [22]. By scanning the database, tables quantifying various correlations were obtained, such as the correlations between two equivalent C_α–C_α distances, or between equivalent main-chain dihedral angles from two related proteins [9]. These relationships were expressed as conditional probability density functions (PDFs) and can be used directly as spatial restraints. For example, probabilities for different values of the main-chain dihedral angles are calculated from the type of residue considered, from main-chain 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 (Fig. 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 [25] 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

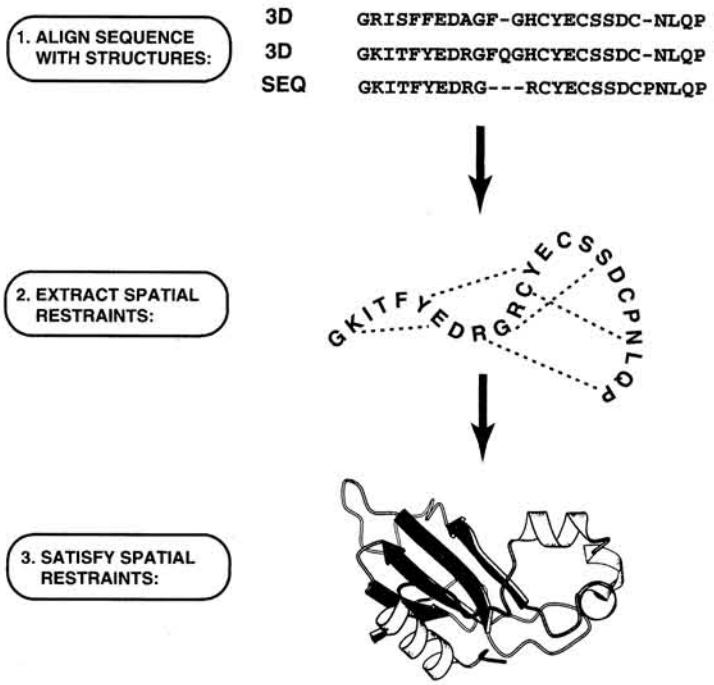

Fig. 1. Comparative protein modeling by satisfaction of spatial restraints. First, the sequence to be modeled (target) is aligned with the known 3D structures (templates). Second, a large number of restraints (broken lines) on distances and dihedral angles in the target sequence are extracted from the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

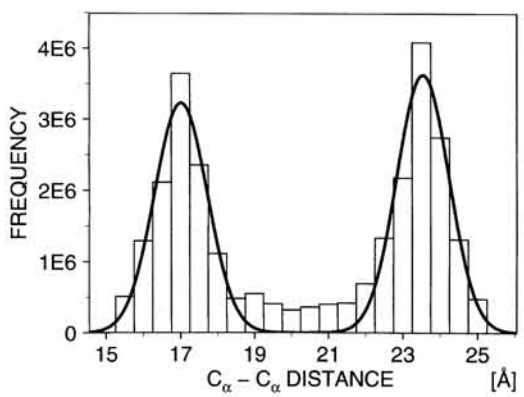

Fig. 2. Sample spatial restraint. A restraint on a given C_α–C_α distance, d, is expressed as a conditional probability density function that depends on two other equivalent distances ($d' = 17.0$ and $d'' = 23.5$): $p(d|d',d'')$. The restraint (continuous line) is obtained by least-squares fitting a 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.

function method [26] employing methods of conjugate gradients and molecular dynamics with simulated annealing [27] (Fig. 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.

3. Errors in homology-derived protein models

In order to facilitate the improvements of comparative modeling by satisfaction of spatial restraints, bona fide predictions of three proteins were made and evaluated when the X-ray structures became available [24]. These three sequences span a range of difficulty from easy (based on 77% sequence identity with templates), and medium (41% sequence identity), to difficult (33% sequence identity).

The three models have good stereochemistry and overall structural accuracy that is as high as the similarity between the template and the actual target structures [24]. The errors that did occur can be

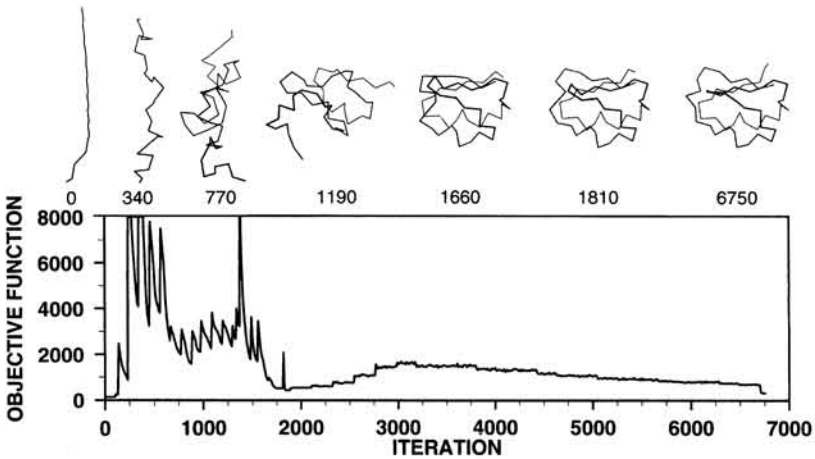

Fig. 3. Optimization of the objective function. This (curve) starts with a random or distorted template structure. The iteration number is indicated below each sample structure. The first 1800 iterations correspond to the variable target function method [26] relying on the conjugate gradients technique. This approach first satisfies sequentially local restraints and slowly introduces longer range restraints until the complete objective function is optimized. In the remaining iterations, molecular dynamics with simulated annealing is used to refine the model [27]. CPU time needed to generate one model is about 20 min for a 250 residue protein on a medium-sized workstation.

divided into four categories: (1) errors in side-chain packing (Fig. 4); (2) distortions or shifts of a region that is aligned correctly with the templates (Fig. 5); (3) distortions or shifts of a region that does not have an equivalent segment in any of the templates (Fig. 6); (4) distortions or shifts of a region that is aligned incorrectly with the templates (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 main-chain atoms are likely to be modeled with an RMS error of about 1 Å. 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 1.5 Å 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

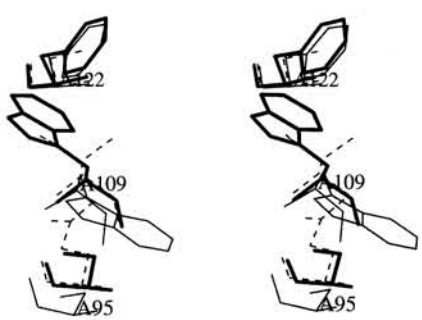

Fig. 4. Errors in side-chain packing. The Trp 109 residue in the crystallographically determined structure of mouse cellular retinoic acid binding protein I [28] (thin line) is compared with its model (thick line), and with the template mouse adipocyte lipid-binding protein (the Brookhaven code 1LIF) (broken line).

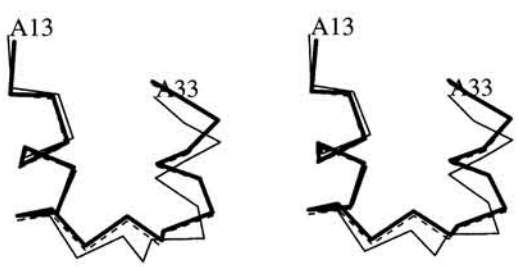

Fig. 5. Distortions and shifts in correctly aligned regions. A region in the crystallographically determined structure of mouse cellular retinoic acid binding protein I [28] (thin line) is compared with its model (thick line), and with the template fatty acid binding protein (the Brookhaven code 2HMB) (broken line).

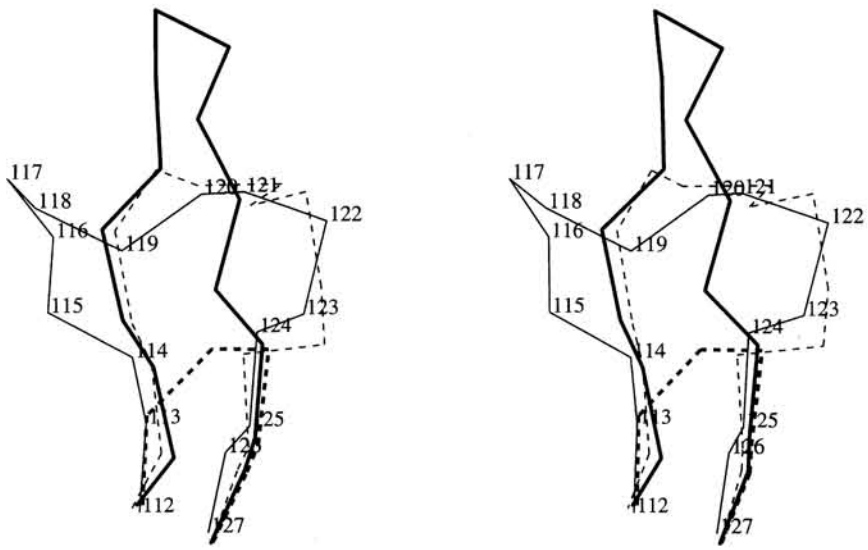

Fig. 6. Errors in unaligned regions. Stereo plot of the C_α trace of the 112–127 loop is shown for the X-ray structure of human eosinophil neurotoxin [29] (continuous thin line), its model (thick line), and the template ribonuclease A structure (residues 111–117; thick broken line).

problems. Insertions longer than about eight residues cannot be modeled accurately at this time, while shorter loops frequently can be modeled successfully [31–35]. 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 3 Å, at this level of sequence similarity [36]. 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 [9,13].

4. Protein structure modeling by satisfaction of spatial restraints

In general, various protein structure modeling problems can be expressed as an optimization of a certain function with respect to the model. The methods differ in the objective function, in the model representation and the degrees of freedom, in

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

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

Fig. 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 5 Å 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 [30].

the method of optimization, and in the starting conformation. The program MODELLER provides much flexibility in all four categories. The 3D model of one or more molecules is obtained by minimizing the objective function F with respect to Cartesian coordinates of up to 15 000 atoms

$$F = F(R) = -\ln\left[\prod_i P_i(f_i/I_i)\right]$$
$$= \sum_i E_i(f_i, a_i) \quad -\ln P_i = E_i \qquad (1)$$

where R are Cartesian coordinates of all atoms, P_i is a conditional probability density function (PDF) for geometric feature f_i that depends on information I_i, E_i is an energy term, and a_i are parameters that generally vary from term to term and are related to I_i. Both the P_i and E_i terms can be seen as spatial restraints. The equivalence between P_i and E_i allows the formulation of a modeling problem that is most convenient for the data at hand: it is possible to use both the ''statistical'' definition of F in terms of PDFs P_i (e.g. when the data come directly from an analysis of the database of known related structures) and the ''physical'' definition of F in terms of the energy terms E_i (e.g. when incorporating the CHARMM force field). Although the statistical and physical definitions are equivalent, it may be easier to arrive at the correct restraint form using one or the other definition. This is illustrated by the following example. The aim is to restrain a certain C_α–C_α distance in a target sequence given two equivalent distances from two template structures. Using the energy terms, a natural restraint would be a sum of two harmonic terms corresponding to each of the template distances. This is equivalent to the PDF consisting of a product of two Gaussian functions. However, it has been shown empirically that the target distance is likely to be close to one of the template distances and less likely to be somewhere in between, described properly as a sum of two Gaussian PDFs (Fig. 2) [9]. The equivalent energy term is thus the negative logarithm of a sum of two Gaussian functions, not the sum of two harmonic terms. In a typical comparative modeling calculation, there are on the order of 40 000 restraints. The form of E_i is simple; it can be a quadratic function, cosine, a logarithm of the weighted sum of a few Gaussian functions, Coulomb law, Lennard-Jones potential, and a cubic spline function. The geometric features presently include a distance, an angle, a dihedral angle, and a pair of dihedral angles between two, three, four, and eight points, respectively. Points correspond to real atoms or to pseudo atoms, such as a gravity center of several real atoms. A pair of dihedral angles can be used to restrain simultaneously such strongly correlated features as the mainchain dihedral angles Φ and Ψ of the same residue. Most terms in the CHARMM energy function are implemented in MODELLER. Molecular representations that correspond to any subset of an all-atom topology library of CHARMM [25] (e.g. all-atom, non-hydrogen atoms, C_α-only) as well as a simplified side-chain model [37] can be used at the present. The optimization is currently carried out by the use of the variable target function method [26] employing methods of conjugate gradients and molecular dynamics with simulated annealing [27]. New functional forms, optimizers, and topologies, such as additional approximate side-chain representations, can be easily incorporated into the program.

5. Discussion

Our future development of comparative modeling will rely on the framework for general objective function definition and optimization provided by the program MODELLER [9], briefly described above. In order to improve comparative modeling in the problematic areas (ie, side-chain packing, distortions in correctly and incorrectly aligned regions, and distortions in unaligned regions), we will rely on this flexible framework to test systematically many possible choices for the objective function, model representation, optimization, and starting conformation. Some of the potential benefits of comparative modeling by satisfaction of spatial restraints are discussed next.

The MODELLER objective function is formally similar to the energy function in molecular mechanics and to the objective function in the NMR refinement. They are all sums of many simple terms, each of which depends on a small number of atoms. However, the general representation of restraints in MODELLER (e.g. cubic splines) will allow an accurate incorporation of many kinds of input information about the structure being modeled. This flexibility will be an important advantage since in many cases the bottleneck in

structure prediction is not the power of the optimizer, but the accuracy of the objective function [9].

Traditionally, comparative modeling consisted of four separate problems: aligning the sequence with related structures, constructing the core of the protein, annealing the loops on the core, and decorating the main-chain with side-chains [1]. Although these three problems are independent from each other in the first approximation, the coupling between the backbone and side-chains, between the conserved core and variable loops, and between the alignment and the model will have to be addressed to improve the accuracy of the final models. For example, if side-chain positions are modeled incorrectly, the backbone of loops may also be predicted incorrectly because the native loop conformation is sometimes stabilized by side-chain–main-chain hydrogen bonds and local electrostatic interactions [31]. Similarly, side-chain conformations depend sensitively on the main-chain conformation [38,39].

Another example of interdependence between currently separately considered aspects of modeling is the coupling between the alignment and the positions of all atoms. If the alignment is incorrect, the atoms will certainly be positioned incorrectly by all the current comparative modeling methods. Even a shift in the alignment by only one residue would produce an RMS error in the backbone atoms on the order of 4 Å. It is estimated that the best methods for aligning sequences with structures align incorrectly about 20% of residues, according to the structure–structure alignments, when the sequence identity between the sequence and structure is about 30% [36]. Likewise, threading methods sometimes fail to produce a correct alignment because of the incorrect placement of residues in space or because of ignoring inserted segments altogether. Comparative modeling by optimization can at least, in principle, treat the coupling between the alignment and the model in one simultaneous optimization of all the aspects of the model and the alignment. The final best model may be found by iteratively changing the alignment and, for each alignment, calculating the best model given the current alignment.

Another strength of comparative modeling by satisfaction of spatial restraints is that constraints or restraints derived from a number of different sources can be added to the homology-derived restraints. For example, restraints could be provided by rules for secondary structure packing [40], analyses of hydrophobicity [41] and correlated mutations [42], empirical potentials of mean force [43], NMR experiments [19], cross-linking experiments, fluorescence spectroscopy, image reconstruction in electron microscopy, site-directed mutagenesis [44], etc. In this way, a homology model, especially in the difficult cases, can be improved by making it consistent with available experimental data and with general knowledge about protein structure.

Acknowledgements

We are grateful to crystallographers R. Williams, G. Klejwegt, A. Jones, S.C. Mosimann, D. Newton, R. Youle, and M.N.G. James for providing the structures for the evaluation of MODELLER before their release to the Brookhaven Protein Databank. We also thank Daša Šali for commenting on the manuscript. R.S. is a Howard Hughes Medical Institute predoctoral fellow.

References

[1] T.L. Blundell, B.L. Sibanda, M.J.E. Sternberg and J.M. Thornton, Nature, 326 (1987) 347–352.
[2] J. Greer, Proteins, 7 (1990) 317–334.
[3] M.S. Johnson, N. Srinivasan, R. Sowdhamini and T.L. Blundell, Crit. Rev. Biochem. Mol. Biol., 29 (1994) 1–68.
[4] J. Bajorath, R. Stenkamp and A. Aruffo, Protein Sci., 2 (1994) 1798–1810.
[5] L. Holm, B. Rost, C. Sander, R. Schneider and G. Vriend, Data Based Modeling of Proteins, in S. Doniach (Ed.), Statistical Mechanics, Protein Structure, and Protein Substrate Interactions, Plenum Press, New York, 1994, p. 277–296.
[6] A. Šali, Curr. Opin. Biotech., 6 (1995) 437–451.
[7] A.M. Lesk and C.H. Chothia, Philos. Trans. R. Soc. London, Ser. B, 317 (1986) 345–356.
[8] M. Levitt, J. Mol. Biol., 226 (1992) 507–533.
[9] A. Šali and T.L. Blundell, J. Mol. Biol., 234 (1993) 779–815.
[10] T.F. Havel and M.E. Snow, J. Mol. Biol., 217 (1991) 1–7.
[11] D.G. George, W.C. Barker and L.T. Hunt, Nucl. Acids Res., 14 (1986) 11–15.
[12] E.E. Abola, F.C. Bernstein, S.H. Bryant, T.F. Koetzle and J. Weng, Protein Data Bank, in F. H. Allen, G. Bergerhoff and R. Sievers (Eds.), Crystallographic Databases — Information, Content, Software Systems, Scientific Applications, Data Commission of the International Union of Crystallography, Bonn, Cambridge, Chester, 1987, p. 107–132.

[13] C.A. Orengo, D.T. Jones and J.M. Thornton, Nature, 372 (1994) 631–634.
[14] A. Šali, R. Matsumoto, H.P. McNeil, M. Karplus and R.L. Stevens, J. Biol. Chem., 268 (1993) 9023–9034.
[15] R. Matsumoto, A. Šali, N. Ghildyal, M. Karplus and R. L. Stevens, J. Biol. Chem., 270 (1995) 19 524–19 531.
[16] A. Caputo, M.N.G. James, J.C. Powers, D. Hudig and R.C. Bleackley, Nature Struct. Biol., 1 (1994) 364–367.
[17] C.S. Ring, E. Sun, J.H. McKerrow, G.K. Lee, P.J. Rosenthal, I.D. Kuntz and F.E. Cohen, Proc. Natl. Acad. Sci. USA, 90 (1993) 3583–3587.
[18] M. Carson, C.E. Bugg, L. Delucas and S. Narayana, Acta Crystallogr., D50 (1994) 889–899.
[19] M.J. Sutcliffe, C.M. Dobson and R.E. Oswald, Biochemistry, 31 (1992) 2962–2970.
[20] A. Šali, J.P. Overington, M.S. Johnson and T.L. Blundell, TIBS, 15 (1990) 235–240.
[21] A. Šali and T.L. Blundell, J. Mol. Biol., 212 (1990) 403–428.
[22] A. Šali and J.P. Overington, Protein Sci., 3 (1994) 1582–1596.
[23] A. Šali and T.L. Blundell, Comparative Protein Modelling by Satisfaction of Spatial Restraints, in H. Bohr and S. Brunak (Eds.), Protein Structure by Distance Analysis, IOS Press, Amsterdam, 1994, pp. 64–86.
[24] A. Šali, L. Potterton, F. Yuan, H. van Vlijmen and M. Karplus, Proteins, 23 (1995) 318–326.
[25] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan and M. Karplus, J. Comp. Chem., 4 (1983) 187–217.
[26] W. Braun and N. Gō, J. Mol. Biol., 186 (1985) 611–626.
[27] G.M. Clore, A.T. Brünger, M. Karplus and A.M. Gronenborn, J. Mol. Biol., 191 (1986) 523–551.
[28] G.J. Kleywegt, T. Bergfors, H. Senn, P. Le Motte, B. Gsell, K. Shudo and T.A. Jones, Structure, 2 (1994) 1241.
[29] S.C. Mosimann, D. Newton, R. Youle and M.N.G. James, in preparation.
[30] W. Kabsch and C. Sander, Biopolymers, 22 (1983) 2577–2637.
[31] C. Mattos, G.A. Petsko and M. Karplus, J. Mol. Biol., 238 (1994) 733–747.
[32] K. Fidelis, P.S. Stern, D. Bacon and J. Moult, Protein Eng., 7 (1994) 953–960.
[33] T.V. Borchert, R.A. Abagyan, K.V.R. Kishan, J.P. Zeelen and R.K. Wierenga, Structure, 1 (1993) 205–213.
[34] V. Collura, J. Higo and J. Garnier, Protein Sci., 2 (1993) 1502–1510.
[35] D. Bassolino-Klimas, R.E. Bruccoleri and S. Subramaniam, Protein Sci., 1 (1992) 1465–1476.
[36] M.S. Johnson and J.P. Overington, J. Mol. Biol., 233 (1993) 716–738.
[37] P. Herzyk and R.E. Hubbard, Proteins, 17 (1993) 310–324.
[38] M.J. McGregor, S.A. Islam and M.J.E. Sternberg, J. Mol. Biol., 198 (1987) 295–310.
[39] R.L. Dunbrack and M. Karplus, Nature Struct. Biol., 1 (1994) 334–340.
[40] F.E. Cohen and I.D. Kuntz, Tertiary Structure Prediction, in G.D. Fasman (Ed.), Prediction of Protein Structure and the Principles of Protein Conformation, Plenum Press, New York, 1989, p. 647–705.
[41] A. Aszódi and W.R. Taylor, Protein Eng., 7 (1994) 633–644.
[42] W.R. Taylor and K. Hatrick, Protein Eng., 7 (1994) 341–348.
[43] M.J. Sippl, J. Mol. Biol., 213 (1990) 859–883.
[44] J.P. Boissel, W.R. Lee, S.R. Presnell, F.E. Cohen and H.F. Bunn, J. Biol. Chem., 268 (1993) 15 983–15 993.