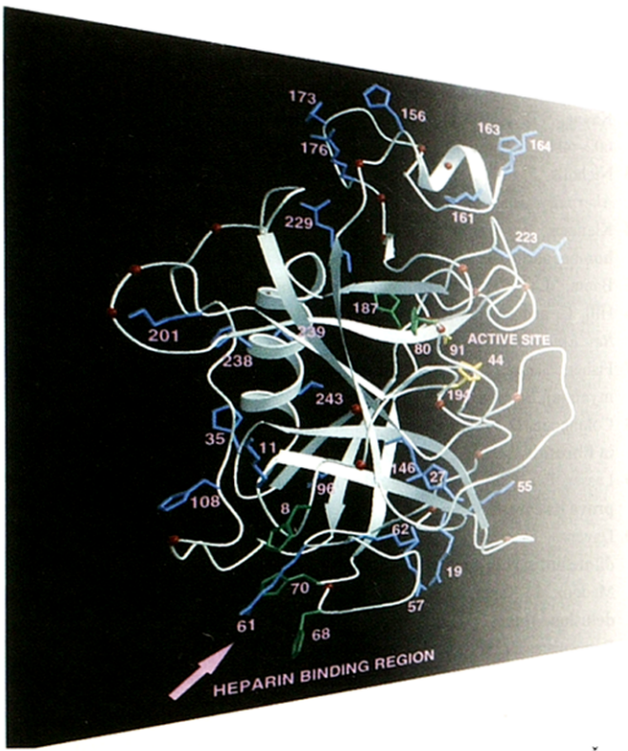
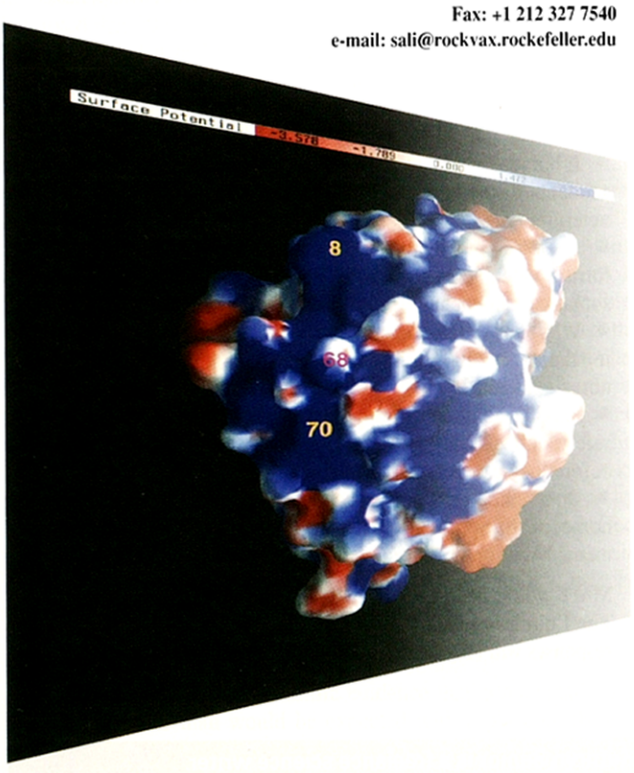


Comparative protein modeling by satisfaction of spatial restraints

Andrej Šali

Andrej Šali
Box 270, The Rockefeller University,
1230 York Avenue, New York,
NY 10021-6399, USA.
Tel: +1 212 327 7550
Fax: +1 212 327 7540
e-mail: sali@rockvax.rockefeller.edu



Approximately one third of known protein sequences are related to at least one known protein structure. As a result, an order of magnitude more sequences can be modeled by comparative modeling than there are experimentally determined protein structures. A large fraction of these models has an accuracy approaching that of a low resolution X-ray structure or a medium resolution nuclear magnetic resonance structure. The number of applications where homology modeling has been proven useful is growing rapidly.

STRUCTURAL biology has demonstrated that knowledge of molecular conformation is a powerful tool in understanding, controlling, and changing the function of biomolecules. While the three-dimensional (3D) structure of proteins can be determined by X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, these experimental techniques are time consuming and not possible if a sufficient quantity of a protein is not available. In addition, NMR spectroscopy cannot yet be applied to proteins larger than ~200 residues, nor to very flexible proteins, while X-ray crystallography depends on the preparation of suitable crystals and on solution of the phase problem, both of which are trial-and-error processes with no guaranteed solutions. By contrast, the amino acid sequence of a protein can be determined relatively easily by molecular biology techniques. As a result, in cases where an experimentally determined structure is not yet available, homology or comparative modeling can frequently predict a useful 3D model of a given sequence (target) by relying on its similarity to proteins with known 3D structure (templates)[1,2] (Fig. 1). This is possible because a small change in the sequence usually results in a small change in the 3D structure[3].

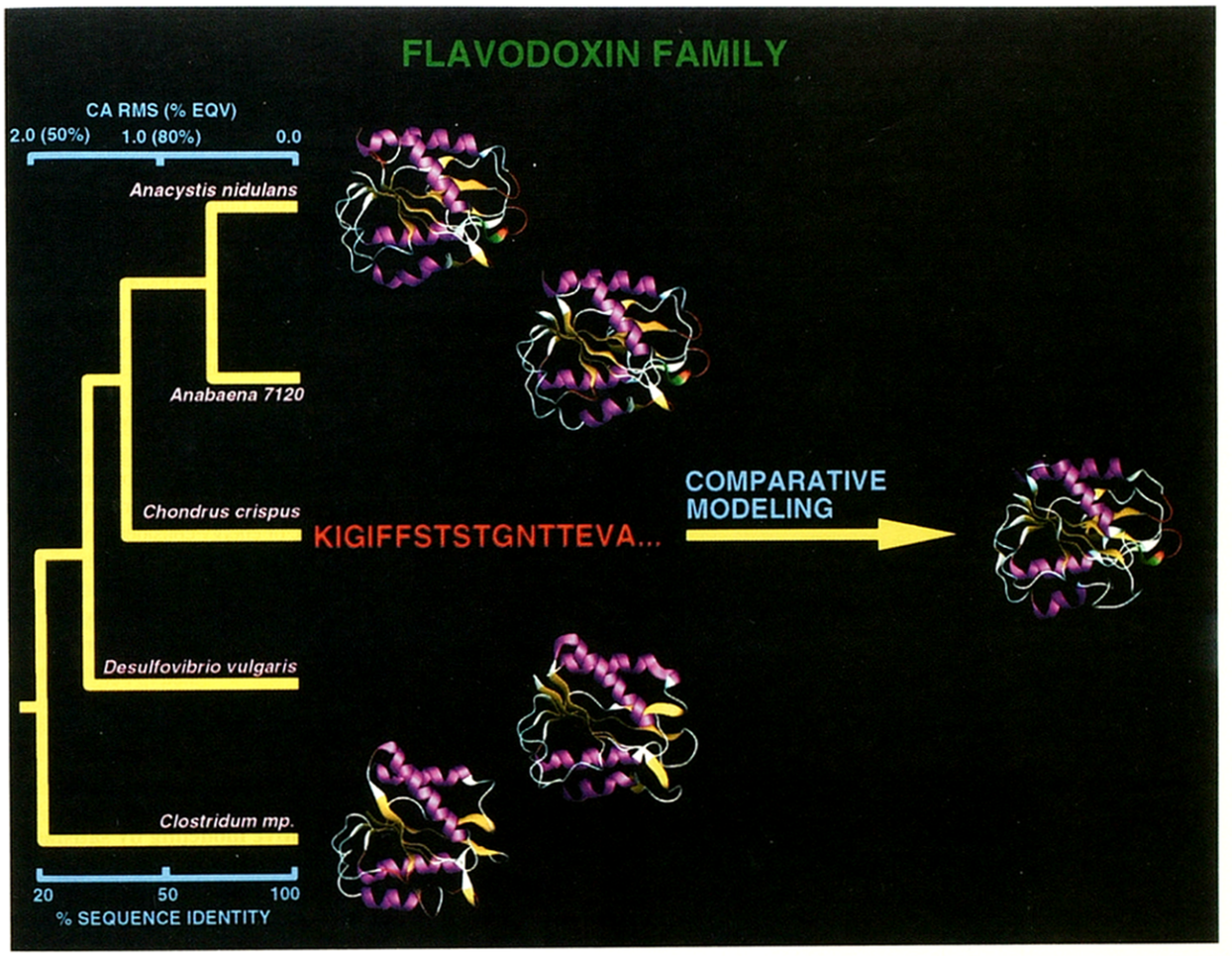


Figure 1. Comparative protein modeling. Comparative modeling is possible because evolution resulted in families of proteins, such as the flavodoxin family, which have both similar sequences and similar three-dimensional (3D) structures. In this example, the 3D structure of a flavodoxin sequence (target) can be modeled using other structures in the same family (templates). The tree shows the sequence similarity (% sequence identity) and structural similarity (the percentage of the C_α atoms that superpose within 3.8 Å of each other and the root-mean-square difference between them) among the members of the family.

Comparative protein modeling

Comparative modeling is useful because about one-third of known sequences appear to be related to at least one known structure[4]. As only ~2000 of the 100 000 known protein sequences have had their structures determined experimentally, the number of sequences that can be modeled relatively accurately 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.

All current comparative-modeling methods consist of four sequential steps (Box 1). 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. The oldest, and still the most widely used, method is modeling by rigid-body assembly[1]. The method constructs the model from a few core regions, and 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 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[10]. 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

Box 1. The stages in comparative protein modeling

(1) Identify proteins with known three-dimensional (3D) structures that are related to the target sequence. One possibility is to use sequence-comparison programs that can match the target sequence with a sequence of each structure in the Brookhaven Protein Data Bank[5]. The sensitivity of the search can be improved if the target sequence is aligned against sequence templates constructed from multiply aligned sequences. Additional sensitivity in detecting remote relationships is gained when structural information about potential homologs is used. Typically, the target sequence is matched against a library of 3D profiles or is threaded through a library of 3D folds[6]. These more sensitive fold-identification techniques are especially useful for finding significant structural relationships when sequence identity drops below 30%.

(2) Align templates with the target sequence. Once all the structures related to the target sequence are identified, a multiple alignment of the target sequence with all the potential template structures is prepared. In principle, most sequence-alignment and structure-comparison methods can be used but, in practice, it is frequently necessary to edit the positions of insertions and deletions manually to ensure that they occur in a reasonable structural context (for example, not in the middle of a helix)[7]. The actual template structures are chosen.

(3) Build the model for the target sequence, given its alignment with the template structures.

(4) Evaluate the model. Models are evaluated using stereochemical criteria[8], 3D profiles[9], and other criteria such as packing and solvent accessibility.

(5) If necessary, repeat the alignment and model building until a satisfactory model is obtained.

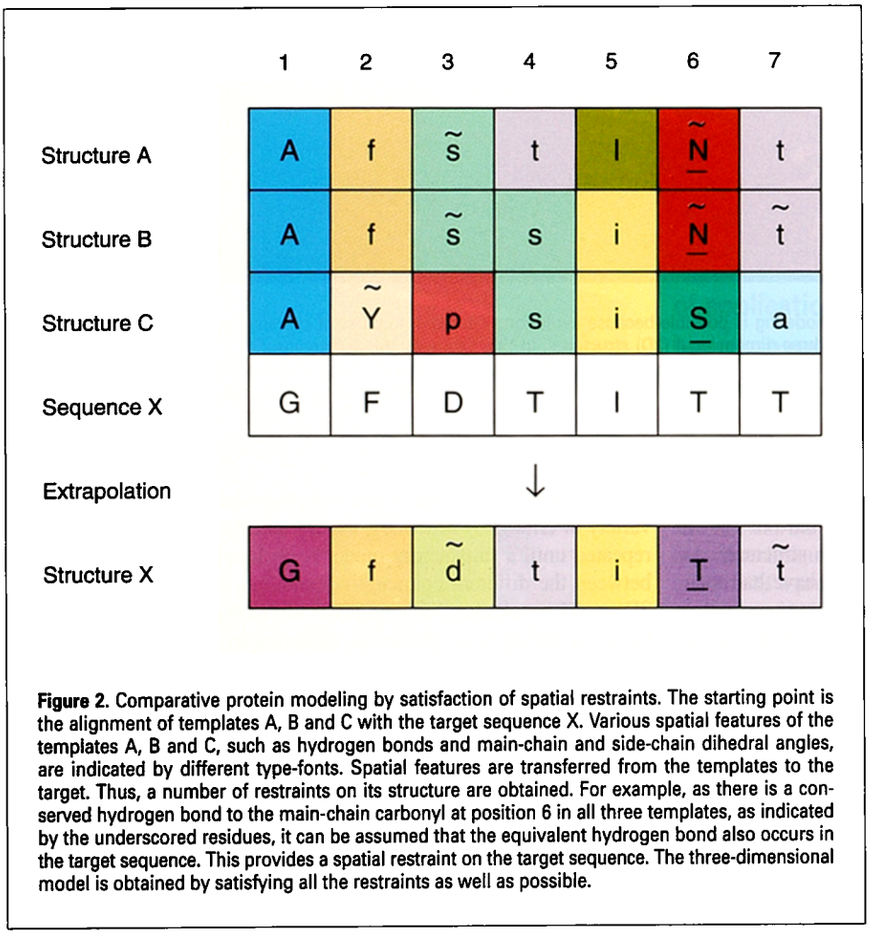


Figure 2. Comparative protein modeling by satisfaction of spatial restraints. The starting point is the alignment of templates A, B and C with the target sequence X. Various spatial features of the templates A, B and C, such as hydrogen bonds and main-chain and side-chain dihedral angles, are indicated by different type-fonts. Spatial features are transferred from the templates to the target. Thus, a number of restraints on its structure are obtained. For example, as there is a conserved hydrogen bond to the main-chain carbonyl at position 6 in all three templates, as indicated by the underscored residues, it can be assumed that the equivalent hydrogen bond also occurs in the target sequence. This provides a spatial restraint on the target sequence. The three-dimensional model is obtained by satisfying all the restraints as well as possible.

restraints, satisfies spatial restraints obtained from the alignment of the target sequence with homologous templates of known structure[11,12]. 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.

In addition to the methods for modeling the whole fold, numerous other techniques for predicting loops and side-chains on a given framework have also been described. These methods can often be used in combination with each other and with comparative modeling techniques. Loops can be calculated by searching a structure database for segments that fit on fixed endpoints, by conformational search with an optional energy minimization, or by a combination of these approaches[13]. As for loops, side-chain conformation has been predicted from similar structures and from steric or energy considerations[14].

This review begins with a description of several recent implementations of restraint-based comparative modeling. Next, the errors in the homology-based models are evaluated and the areas that need improvement are pointed out. Finally, several applications of comparative modeling are described to illustrate the types of problems that can be solved by this technique.

Comparative modeling by satisfaction of spatial restraints

All restraint-based approaches to comparative protein modeling extract distance and/or dihedral angle restraints on the target sequence from the alignment with related structures, add restraints implied by the covalent topology (stereochemical restraints), and calculate the model by minimizing the violations of all restraints[12,15–18] (Fig. 2). Thus, the two main differences between the various approaches are in the derivation and satisfaction of spatial restraints. In this section, each of the recently published approaches is described.

In the method by Havel and Snow[12,19], lower and upper bounds on C_α–C_α and main-chain–side-chain distances, hydrogen bonds, and conserved dihedral angles were derived for *Escherichia coli* flavodoxin from four other flavodoxins; bounds were calculated for all distances and dihedral angles that had equivalent atoms in the template structures. The allowed range of values of a distance or a dihedral angle depended on the degree of structural variability at the corresponding position in the template structures. Distance geometry was used to obtain an ensemble of

Glossary

Brookhaven Protein Data Bank – A collection of protein three-dimensional structures determined by X-ray crystallography or nuclear magnetic resonance spectroscopy. It is accessible at the WWW address http://www.pdb.bnl.gov/.

Conditional probability density function – A function that specifies the probability for each possible value of the restrained feature (for example, the distance between two atoms), given some known features (for example, an equivalent distance in a related structure). This is the most general mathematical formulation of a spatial restraint.

Constraints and restraints – Constraints restrict a spatial feature, such as a distance between two atoms, to a particular single value. Restraints allow a wider range of values, possibly with varying probabilities.

Dihedral angle – An angle formed by four points *i, j, k, l* (for example, atoms). It is defined as the angle between the normals of planes *ijk* and *jkl*.

Distance geometry – A technique for calculating Cartesian coordinates of points from lower and upper bounds on some distances between these points.

Energy minimization – A technique that changes the conformation of a molecule in order to decrease its energy as much as possible.

Lennard-Jones potential – An energy term that is frequently used to describe an interaction between a pair of particles: $E = A/d^{12} - B/d^6$, where A and B are positive constants and d is the distance between the particles.

R factor – A measure used in X-ray crystallography to quantify the difference between the experimentally observed structure factors and those calculated from the current model. Typical values are 40%, 25% and 15% for unrefined, medium and highly refined structures, respectively.

RMS difference – Root-mean-square difference measures structural dissimilarity between two superimposed structures. It is defined as $\sqrt{[(\sum_i^N d_i^2)/N]}$, where the sum runs over N equivalent pairs of atoms, one from each structure, and d_i is the distance between the two atoms in the i-th pair.

Simulated annealing with molecular dynamics – A powerful stochastic optimization technique that involves simulating Newtonian equations of motion at increasingly lower temperatures, pretending that the objective function corresponds to the energy function.

Stereochemical restraints – These spatial restraints are implied by the covalent topology of the molecule. They include restraints on bond lengths, bond angles, dihedral angles, planarity of rings, chirality of chiral centers and on non-bonded atom–atom overlaps.

Threading – A sensitive method for detecting remote sequence–structure relationships and for aligning a sequence with a structure. Structural information and potentials of mean force are used.

Variable target function method – A deterministic optimization technique that involves conjugate gradients optimization of an increasingly more complicated objective function, starting with an easy to optimize subset of restraints and ending with the full objective function, including all restraints.

approximate 3D models, which were then exhaustively refined by restrained molecular dynamics with simulated annealing in water.

Comparative modeling by optimization of a potential function constructed from a sequence alignment with related structures was described by Snow[16]. A model consisted of C_α atoms that were restrained by a form of a Lennard-Jones potential. The position of the minimum of each Lennard-Jones term corresponded to a weighted average of equivalent distances in homologous structures, and the depth of the minimum was inversely proportional to the variability among these distances. The 'energy' was minimized by a simulated annealing procedure in the angle and dihedral-angle space, followed by a conjugate gradients refinement of the atomic positions. The method was tested by modeling rubredoxin on the basis of four other rubredoxin structures.

The method developed by Srinivasan *et al.*[18] used a single template structure to obtain distance constraints on the target sequence. As in Ref. 12, constraints were derived for all pairs of atoms that had equivalent pairs in the template structure. Distance constraints were satisfied by a distance-geometry program and a subsequent energy refinement. When the template and target sequences are similar, the target model is also very similar to the template structure. Subsequently, the method was improved by relaxing distance constraints on the target sequence outside the manually delineated structurally conserved regions[20]. This relaxation facilitated 3D embedding and energy minimization, and increased the root-mean-square (RMS) difference between the template and the model, but it does not appear to increase the accuracy of the model beyond the similarity between the template and the actual structure of the target[20].

Brocklehurst and Perham described an automated method for constructing a 3D model of a sequence that is aligned with related structures[17]. This method optimizes a relatively small number of spatial restraints that are judged to be important for the fold and/or function, and thus more likely to be conserved in the family of proteins. These restraints restrain main-chain hydrogen bonds, attractive van der Waals contacts, and main-chain and side-chain dihedral angles. The optimization consists of molecular dynamics with simulated annealing. The method has been applied to two domains from the dehydrogenase family.

Our own approach is now described in more detail[11,15,21,22]. The question addressed is: what is the most probable structure for a

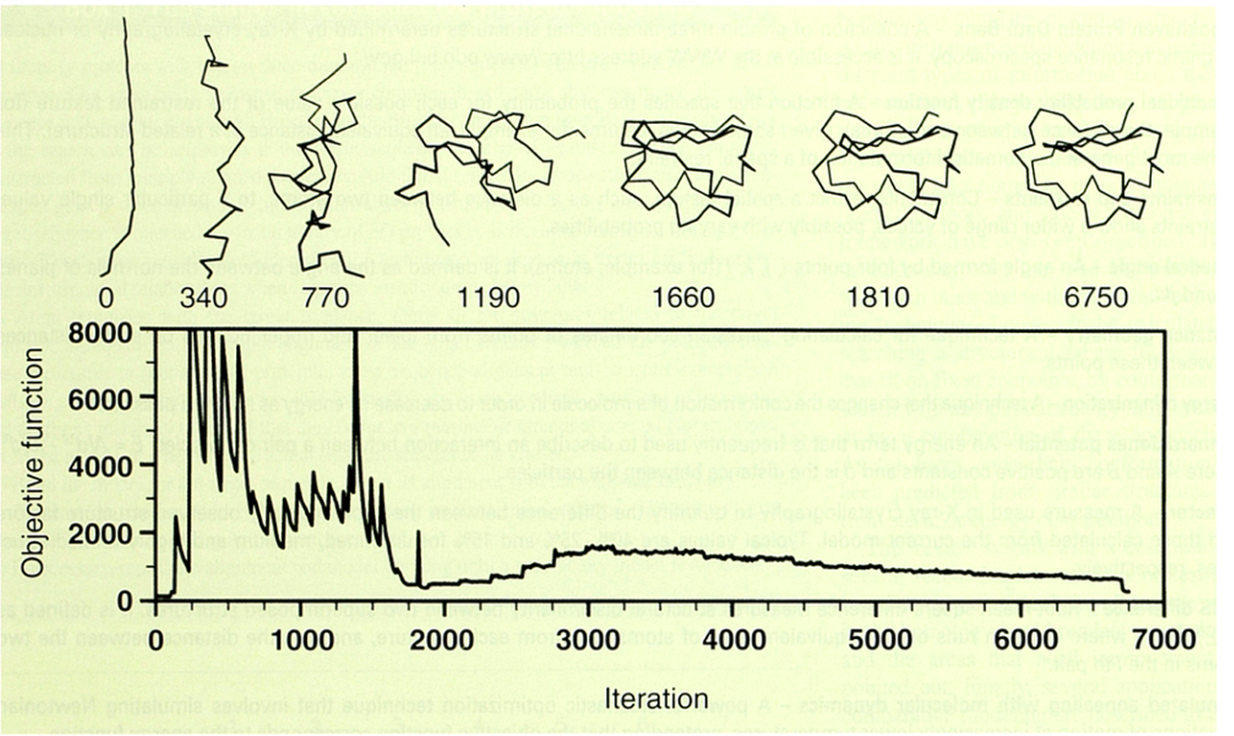


Figure 3. Optimization of the objective function by MODELLER. Optimization of the objective function (curve) starts with a random structure or a distorted template structure. The iteration number is indicated below each sample structure. The first ~2000 iterations correspond to the variable target function method[23], 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 4750 iterations, molecular dynamics with simulated annealing is used to refine the model, using all homology and stereochemical restraints.

certain sequence given its alignment with related structures? It is implemented in the computer program MODELLER*. The input to the program is 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 heavy atoms. First, MODELLER derives many distance and dihedral-angle restraints on the target sequence from its alignment with template 3D structures. Spatial restraints on the target sequence are obtained from the statistical analysis of the relationships between various features of protein structure. A database of 105 family alignments, including 416 proteins with known 3D structure, was constructed[21] to obtain the tables quantifying the relationships, such as those between two equivalent C_α–C_α distances, or between equivalent main-chain dihedral angles from two related proteins. These relationships were expressed as conditional probability density distributions 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. An important difference from the other methods discussed in this section is that the spatial restraints are obtained empirically from a database and are not guessed. Next, the homology-derived restraints and energy terms enforcing proper stereochemistry are combined into an objective function. Finally, the model is obtained by optimizing the objective function in Cartesian space. This optimization is carried out by the use of the variable target function method[23], employing methods of conjugate gradients and molecular dynamics with simulated annealing (Fig. 3). Several slightly different models can be calculated by varying the initial structure.

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

Unification of modeling and refinement techniques

One of the strengths 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[24], analyses of hydrophobicity[25] and correlated mutations[26], empirical potentials of mean force[27], NMR experiments[28], cross-linking experiments, fluorescence spectroscopy, image reconstruction in electron microscopy, site-directed mutagenesis[29], etc. In this way, a homology model, particularly in the difficult cases, could be improved by making it consistent with available experimental data and/or with more general knowledge about protein structure. In

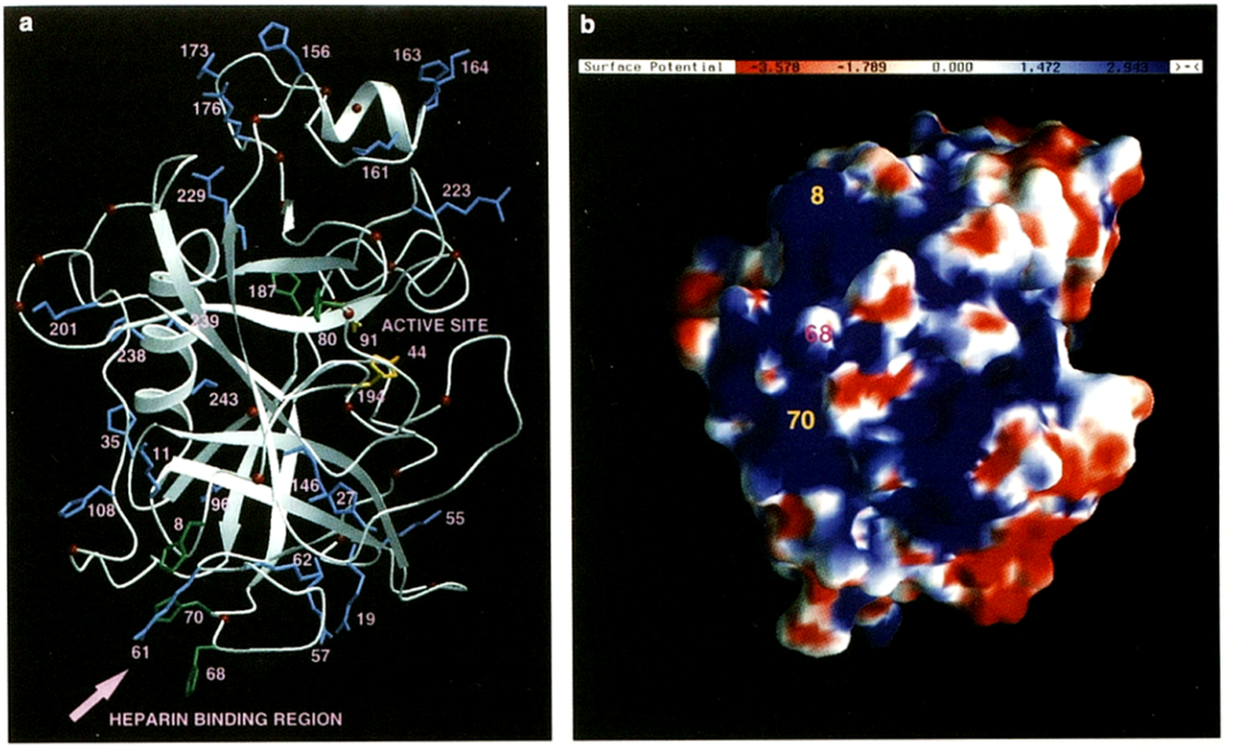


Figure 4. Comparative protein modeling of mouse mast cell protease 7 (mMCP-7). (a) Ribbon diagram of the three-dimensional model of mMCP-7. The positively charged side chains (Lys, Arg and His), the C_α atoms of the negatively charged residues (Asp, Glu), the active site side chains (His44, Asp91 and Ser194) and the five mutated His side-chains are shown in blue, red, black and green, respectively. The heparin-binding region, which contains many positively charged residues and few negatively charged residues, is indicated by an arrow. The figure was prepared by program MOLSCRIPT (Ref. 39). (b) Electrostatic potential at the positively charged region of mMCP-7, at the acidic pH of the granule, as calculated by program GRASP (Ref. 40). The molecular surface is colored by the electrostatic potential, as shown by the color bar on the panel (in units of kT). The view is obtained by rotating the molecule in panel (a) by 90° around a horizontal axis. The three His residues in the positively charged region, which were proven to be important for heparin binding by site-directed mutagenesis, are numbered.

fact, every protein-modeling method that is based on an optimization of a certain function can be seen as restraint-based modeling. For example, in energy minimization and molecular-dynamics methods, the restraints correspond with the energy terms; in NMR structure determination, restraints correspond with experimentally determined lower and upper bounds on distances and dihedral angles; in X-ray crystallography, the restraints correspond with the structure factors; and in folding simulations using simplified models, restraints correspond with the potential of mean force.

Errors in models

Comparative modeling remains the only modeling method that can provide models with an RMS error lower than 2 Å. In general, the best restraint-based comparative models have good stereochemistry and overall structural accuracy that is at least as high as the similarity between the template and the actual target structure[22]. When more than one template is available, the model may be closer to the actual structure than any of the templates[11].

When the target sequence is at least 40% identical to one or more of the templates, the model can have the main-chain RMS error as low as 1 Å for 90% of the residues[22]. In this range of sequence similarity, the alignment is straightforward to construct, there are not many gaps, and structural differences between the proteins are usually limited to loops and side-chains. As a result, modeling can be done in an automated way without any significant loss of accuracy relative to careful model building by an expert[22].

When sequence identity between the target and the templates is between 30% and 40%, structural differences become larger, and gaps in the alignment are more frequent and longer. As a result, the main-chain RMS error rises to ~1.5 Å for ~80% of residues[22]. The rest of the residues are modeled with large errors because the methods generally cannot model structural distortions, rigid-body shifts and insertions, and cannot recover from misalignments. In particular, insertions longer than about eight residues cannot be modeled accurately at this time, even when the alignment of the stem regions delimiting the insertion is correct; many of the insertions shorter than eight residues also cannot be modeled successfully, primarily because the alignment of the inserted and neighboring residues is frequently incorrect. If the length of an insertion can be extended sufficiently to make the alignment of the delimiting stem regions reliable, but not too much, so that less than eight residues are inserted, the insertions can frequently be modeled successfully[13]. Below 40%

sequence identity, large errors in the alignment can sometimes be prevented by examining and editing the alignment manually. In general, it can be expected that about 20% of residues will be misaligned and, consequently, incorrectly modeled, when the sequence identity between the target and the templates is 30% (Ref. 30). When sequence identity drops below 30%, the main problem becomes the identification of related templates and their alignment with the sequence to be modeled[6]. Model-evaluation methods can be used to identify the inaccurately modeled regions of a protein (Box 1), and thus to determine the usefulness of a given model for a particular application.

To put the errors in comparative modeling into perspective, the differences among experimentally determined structures of the same protein are listed. The 1 Å accuracy of main-chain atom positions corresponds with X-ray structures defined at a resolution of ~2.5 Å and with an R factor of ~25% (Ref. 31), as well as with NMR structures determined from ten interproton distance restraints per residue[32]. Similarly, differences between highly refined X-ray and NMR structures of the same protein also tend to be ~1 Å (Ref. 32). Changes in the environment (for example, crystal packing, solvent, ligands) can also have a significant effect on the structure. Overall, homology 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. 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[33,34].

Applications

Comparative modeling, even if less accurate than high-resolution experimental methods, can be helpful in proposing and testing hypotheses in molecular biology. Some experimentally validated studies include the design of micromolar inhibitors of the malarial cysteine protease[35], the prediction and conversion of substrate specificity of granzyme B (Ref. 36), the solution of a molecular replacement problem in X-ray crystallography[37], and providing starting models for the refinement in the NMR spectroscopy[28]. With additional developments, the homology-derived restraints might also help guide the X-ray and NMR refinement process itself, in a similar way to the bias exerted by the molecular mechanics force fields that are currently in use.

To illustrate the role of comparative modeling in molecular biology, one application is described in more detail. Mouse mast cell protease[7] (mMCP-7) is a tryptase stored in the secretory granules of mast cells. At the granule pH of 5.5, mMCP-7 is fully active and is bound to heparin-containing serglycin proteoglycans. To understand the interaction between mMCP-7 and heparin inside and outside the mast cell, this tryptase was first studied by comparative protein modeling[33,34] (Fig. 4a), and later by site-directed mutagenesis[34]. Although mMCP-7 lacks known linear sequences of amino acids that interact with heparin, the 3D model of mMCP-7 revealed an area on the surface of the folded protein that had many Lys, Arg and His residues, and few Asp and Glu residues. This area exhibits a strong positive electrostatic potential at the acidic pH of the granule, but not at neutral pH (Fig. 4b). In agreement with this calculation, recombinant pro-mMCP-7 bound to a heparin-affinity column at pH 5.5 and readily dissociated from the column at pH > 6.5. Site-directed mutagenesis confirmed the prediction that the conversion of His residues 8, 68 and 70 in the positively charged region into Glu residues prevents the binding of pro-mMCP-7 to heparin. It would have been much more time consuming and expensive to identify the binding site without the modeling.

Conclusions

The number of sequences that can be modeled by comparative methods is an order of magnitude larger than the number of experimentally determined protein structures. On one hand, the accuracy of a large fraction of these models is, in many ways, comparable with the accuracy of low resolution X-ray structures and medium resolution NMR structures. On the other hand, homology-based models may have large errors in some regions. In particular, it is still not possible to model accurately distortions and shifts in structure relative to templates and insertions longer than about eight residues, even when the alignment is correct. In addition, many insertions that are shorter than eight residues cannot be modeled successfully, primarily because of mistakes in alignment.

Future improvements in comparative modeling should aim to model proteins with lower similarities to known structures, to increase the accuracy of the models, and to make modeling fully automated. The improvements are likely to include the simultaneous optimization of side-chain and backbone conformations in side-chain modeling, and simultaneous optimization of a loop and its environment in loop modeling. 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. The addition of atomic or residue-based potentials of mean force to the homology-derived scoring function, such as that of MODELLER (Ref. 11), could be one way of achieving this goal. 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[38], are needed. An automation of this process will be very useful because at least one half of all related protein pairs are related at less than 40% sequence identity[11].

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 even more useful in the future.

The outstanding questions

- How can more proteins be modeled in a more accurate and automated fashion?
- What is the best alignment for comparative modeling of a sequence that is only remotely related to template structures?
- How can side-chains be modeled given an approximate backbone?
- How can insertions be modeled, particularly those that are longer than eight residues?
- What potentials of mean force are best for modeling distortions and rigid-body shifts?

Acknowledgements. I am grateful to John Overington, Tom Blundell, Martin Karplus and Mark Johnson for discussions concerning comparative protein modeling. I would also like to thank Daša Šali for her comments on the manuscript.

References

1 Johnson, M., Srinivasan, N., Sowdhamini, R. and Blundell, T. (1994) **Knowledge-based protein modelling,** *CRC Crit. Rev. Biochem. Mol. Biol.* 29, 1–68

2 Šali, A. (1995) **Modeling mutations and homologous proteins,** *Curr. Opin. Biotechnol.* 6, 437–451

3 Lesk, A. and Chothia, C. (1986) **The response of protein structures to amino acid sequence changes,** *Philos. Trans. R. Soc. London Ser. B* 317, 345–356

4 Orengo, C., Jones, D. and Thornton, J. (1994) **Protein superfamilies and domain superfolds,** *Nature* 372, 631–634

5 Abola, E. *et al.* (1987) **Protein data bank,** in *Crystallographic Databases – Information, Content, Software Systems, Scientific Applications* (Allen, F., Bergerhoff, G. and Sievers, R., eds), pp. 107–132, Data Commission of the International Union of Crystallography

6 Johnson, M. (1995) **Cornering and catching the common protein fold,** *Mol. Med. Today* 1, 188–194

7 Holm, L. and Sander, C. (1994) **Searching protein structure databases has come of age,** *Proteins* 19, 165–173

8 Laskowski, R., McArthur, M., Moss, D. and Thornton, J. (1993) **PROCHECK: A program to check the stereochemical quality of protein structures,** *J. Appl. Crystallogr.* 26, 283–291

9 Wodak, S. and Rooman, M. (1993) **Generating and testing protein folds,** *Curr. Opin. Struct. Biol.* 3, 247–259

10 Levitt, M. (1992) **Accurate modeling of protein conformation by automatic segment matching,** *J. Mol. Biol.* 226, 507–533

11 Šali, A. and Blundell, T. (1993) **Comparative protein modelling by satisfaction of spatial restraints,** *J. Mol. Biol.* 234, 779–815

12 Havel, T. and Snow, M. (1991) **A new method for building protein conformations from sequence alignments with homologue of known structure,** *J. Mol. Biol.* 217, 1–7

13 Fidelis, K., Stern, P., Bacon, D. and Moult, J. (1994) **Comparison of systematic search and database methods for constructing segments of protein structure,** *Protein Eng.* 7, 953–960

14 Lee, C. (1994) **Predicting protein mutant energetics by self consistent ensemble optimisation,** *J. Mol. Biol.* 236, 918–939

15 Šali, A., Overington, J., Johnson, M. and Blundell, T. (1990) **From comparisons of protein sequences and structures to protein modelling and design,** *Trends Biochem. Sci.* 15, 235–240

16 Snow, M. (1993) **A novel parameterization scheme for energy equations and its use to calculate the structure of protein molecules,** *Proteins* 15, 183–190

17 Brocklehurst, S. and Perham, R. (1993) **Prediction of the three-dimensional structures of the biotinylated domain from yeast pyruvate carboxylase and of the lipolyated H-protein from the pea leaf glycine cleavage system: A new automated method for the prediction of protein tertiary structure,** *Protein Sci.* 2, 626–639

18 Srinivasan, S., March, C. and Sudarsanam, S. (1993) **An automated method for modeling proteins on known templates using distance geometry,** *Protein Sci.* 2, 227–289

19 Havel, T. (1993) **Predicting the structure of the flavodoxin from *Escherichia coli* by homology modeling, distance geometry and molecular dynamicism,** *Mol. Simulation* 10, 175–210

20 Sudarsanam, S., March, C. and Srinivasan, S. (1994) **Homology modeling of divergent proteins,** *J. Mol. Biol.* 241, 143–149

21 Šali, A. and Overington, J. (1994) **Derivation of rules for comparative protein modeling from a database of protein structure alignments,** *Protein Sci.* 3, 1582–1596

22 Šali, A. *et al.* **Evaluation of comparative protein modeling by MODELLER,** *Proteins* (in press)

23 Braun, W. and Gō, N. (1985) **Calculation of protein conformations by proton–proton distance constraints: A new efficient algorithm,** *J. Mol. Biol.* 186, 611–626

24 Cohen, F. and Kuntz, I. (1989) **Tertiary structure prediction,** in *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G., ed.), pp. 647–705, Plenum Press

25 Aszódi, A. and Taylor, W. (1994) **Secondary structure formation in model polypeptide chains,** *Protein Eng.* 7, 633–644

26 Taylor, W. and Hatrick, K. (1994) **Compensating changes in protein multiple sequence alignments,** *Protein Eng.* 7, 341–348

27 Sippl, M. (1990) **Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins,** *J. Mol. Biol.* 213, 859–883

28 Sutcliffe, M., Dobson, C. and Oswald, R. (1992) **Solution structure of neuronal bungarotoxin determined by two-dimensional NMR spectroscopy: Calculation of tertiary structure using systematic homologous model building, dynamical simulated annealing, and restrained molecular dynamics,** *Biochemistry* 31, 2962–2970

29 Boissel, J. *et al.* (1993) **Erythropoietin structure–function relationships. Mutant proteins that test a model of tertiary structure,** *J. Biol. Chem.* 268, 15983–15993

30 Johnson, M. and Overington, J. (1993) **A structural basis for sequence comparisons: An evaluation of scoring methodologies,** *J. Mol. Biol.* 233, 716–738

31 Ohlendorf, D. (1994) **Accuracy of refined protein structures. II Comparison of four independently refined models of human interleukin 1β,** *Acta Crystallogr.* D50, 808–812

32 Clore, G., Robien, M. and Gronenborn, A. (1993) **Exploring the limits of precision and accuracy of protein structures determined by nuclear magnetic resonance spectroscopy,** *J. Mol. Biol.* 231, 82–102

33 Šali, A. *et al.* (1993) **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

34 Matsumoto, R. *et al.* (1995) **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

35 Ring, C. *et al.* (1993) **Structure-based inhibitor design by using protein models for the development of antiparasitic agents,** *Proc. Natl Acad. Sci. USA* 90, 3583–3587

36 Caputo, A. *et al.* (1994) **Conversion of the substrate specificity of mouse proteinase granzyme B,** *Nature Struct. Biol.* 1, 364–367

37 Carson, M., Bugg, C., Delucas, L. and Narayana, S. (1994) **Comparison of homology models with the experimental structure of a novel serine protease,** *Acta Crystallogr.* D50, 889–899

38 Jones, D., Taylor, W. and Thornton, J. (1992) **A new approach to protein fold recognition,** *Nature* 358, 86–89

39 Kraulis, P. (1991) **MOLSCRIPT: A program to produce both detailed and schematic plots of protein structure,** *J. Appl. Crystallogr.* 24, 946–950

40 Nicholls, A., Sharp, K. and Honig, B. (1991) **Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons,** *Proteins* 11, 281–296