Advances in comparative protein-structure modelling

Roberto Sánchez and Andrej Šali*

Comparative modelling of protein 3D structure can now be applied with reasonable accuracy to ten times more protein sequences than the number of experimentally determined protein structures. A protein sequence that has at least 40% identity to a known structure can be modelled automatically with an accuracy approaching that of a low resolution X-ray structure or a medium resolution NMR structure. Currently, the errors in comparative models include mistakes in the packing of sidechains, in the conformation and shifts of the core segments and loops, and, most importantly, in an incorrect alignment of the modelled sequence with related known structures. Nevertheless, the number of applications in which comparative modelling has been proven to be useful has grown rapidly.

Addresses
Box 270, The Rockefeller University, 1230 York Avenue, New York, NY 10021-6399, USA
*e-mail: sali@rockvax.rockefeller.edu

Current Opinion in Structural Biology 1997, 7:206–214

Electronic identifier: 0959-440X-007-00206

Abbreviations

3D	three-dimensional
MD	molecular dynamics
RMS	root mean square
RMSD	RMS deviation

Introduction

Comparative or homology protein modelling uses experimentally determined protein structures (templates) to predict the conformation of another protein that has a similar amino acid sequence (the target) [1–3,4•]. This approach to modelling is possible because a small change in the protein sequence usually results in a small change in its 3D structure [5,6]. Comparative modelling remains the only modelling method that can provide models with a root mean square (rms) error lower than 2 Å.

All current comparative-modelling methods consist of four sequential steps [3]. 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 the 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-modelling 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 modelling using rigid-body assembly [7–9]. This method constructs the model from a few core regions and from loops and sidechains, which are obtained from dissecting related structures. The assembly involves fitting the rigid bodies on the framework, which is defined as the average of the Cα atoms in the conserved regions of the fold. Another family of methods, modelling by segment matching, relies on the approximate positions of conserved atoms from the templates to calculate the coordinates of other atoms [10–13]. This is achieved using a database of short segments of protein structure, energy or geometry rules, or some combination of these criteria. The third group of methods, modelling by satisfaction of spatial restraints, uses either distance geometry [14,15] or optimization techniques [16] to satisfy spatial restraints obtained from the alignment of the target sequence with homologous templates of known structure. In addition to the methods for modelling the whole fold, numerous other techniques for predicting loops [17] and sidechains [18•] on a given backbone have also been described. These methods can often be used in combination with each other and with comparative-modelling techniques.

In this review, we discuss papers on the modelling of loops, sidechains and whole protein structures that have been published in the past year. In addition, we evaluate the accuracy of comparative models and discuss the role of comparative modelling in making full use of protein sequences in genome databases. We conclude with comments on the future challenges for comparative modellers.

Modelling of loops

Loops can be calculated by searching a protein-structure database for segments that fit on fixed-backbone endpoints [10], by using a conformational search with an optional energy minimization [19–21], or by a combination of these two methods [22,23]. Many different implementations of the three approaches have been proposed (see review [3] and [24•,25,26••]).

Zheng and Kyle [26••] describe a new version of their loop-modelling method [27]. The random starting conformation for optimization has all its bonds scaled so that the loop fits onto the anchor regions. The loop is then relaxed to its standard geometry in the protein environment by energy minimization using the CHARMm forcefield [28]. The method is combined with multiple

copy sampling to increase its efficiency up to a factor of five. The sampling is biased to more populated regions of the (ψ,ϕ) map. Zheng and Kyle [26••] have calculated conformations of eight seven-residue loops embedded in the correct environment of the crystallographic protein structures. The main improvement in their method is that loop conformations are sampled more extensively. This decreases the average rms error for the backbone atoms from 1.1 Å to 0.7 Å. The accuracy of the loop models obtained in this study can probably not be expected in real modelling situations, even when the alignment is correct. The reason for this is that the correct environment of the loop, which can act as a mould, is not known in typical comparative-modelling problems.

Donate *et al.* [29••] describe an analysis of loops that is likely to be helpful in future prediction efforts. 2 024 loops of one to eight residues in length have been identified. These loops, grouped according to their length and type of their bounding secondary-structure segments, have been superimposed and clustered into 161 conformational classes, covering 63% of all loops. The relative disposition of the bounding secondary-structure elements varies among the classes. For each class, amino acid type preferences of some positions have been identified and expressed in terms of key residues. Most of these residues have been involved in stabilizing loop conformation, often via a positive ϕ conformation or a secondary-structure capping. The database can be used in loop modelling by comparing the sequence of the loop to be modelled and the spatial disposition of its anchoring secondary-structure segments with the potential template loops in the database.

Kwasigroch *et al.* [30•] describe a similar study of loops that is also likely to be useful in loop modelling. A database containing loops of three to eight residues long has been built. Loops have been divided into two parts: the side residues that directly bond to the flanking secondary-structure segments; and the inner section. The conformations of the side residues have been found to be correlated to those of the flanking secondary-structure segments, whereas the inner residues adopt conformations uncorrelated from one residue to the next. Loops of the same length are clustered into families of loops having similar conformations. For each cluster, residue positions are determined that have a nonrandom residue type and/or conformation. Despite the conserved conformation of the inner part of the loops within each cluster, the loop termini can show a high degree of structural variability.

Modelling of sidechains

As for loops, sidechain conformation has been predicted from similar structures, from proteins in general, and from steric or energy considerations. Many different implementations of these approaches have been proposed (see reviews [3,18•] and [31–37,38••–42••,43]).

Dunbrack *et al.* [38••] recalculate and extensively evaluate their mainchain dependent sidechain rotamer library. Their library gives the probabilities for sidechain rotamers that depend on the mainchain (ψ, ϕ) values as well as the residue type; it is available on Internet [44]. The multivariate rotamer library is justified by the significant correlations between the sidechain dihedral angles and the backbone (ψ,ϕ) values. The initial sidechain conformations for optimization on a fixed backbone have been obtained according to the rotamer library and the conformation of the equivalent template sidechains. The subsequent combinatorial optimization is designed to remove most steric clashes. The accuracy of the method reaches 82% for χ_1 dihedral angles and 72% for both χ_1 and χ_2 dihedral angles when the backbones from the templates in the range of 30–90% sequence identity are used; a prediction is deemed correct when within 40° of the target crystal structure value. This appears to be one of the most accurate methods for sidechain prediction.

Lee [39••] evaluates the accuracy of structural and thermodynamic predictions using his sidechain-modelling method. The main purpose is to measure the errors caused by the fixed-backbone approximation. Sidechain conformations of several single mutants of T4 lysozyme have been modelled on the wild-type backbone. The method builds sidechains on a fixed backbone by relying on the self-consistent mean field approximation [45,46]. The energetics are described by only Lennard–Jones and simple dihedral-angle terms. Two schemes for sidechain dihedral angles have been explored: discrete rotamers (the 'rotamer model'); and more flexible sidechains that have approximately 10° bins for χ_i values (the 'continuous model'). The rotamer model is more affected by the backbone shifts than the continuous model, which is able to accommodate the sidechains on the wild-type backbone by using distorted sidechain torsion angles. The predicted stability of the mutants using the 'continuous model' shows a good correlation with experimental values. Mainchain shifts of up to 0.5 Å cause increased sidechain coordinate errors of up to 0.8 Å, torsional errors of 10–30°, and exaggerated strain energy for overpacked mutants, compared with the same calculations performed with the correct mutant backbones.

Scheraga and coworkers [40••] use a set of known protein structures to derive continuous Gaussian trivariate (ψ,ϕ,χ_1) and bivariate (ψ,ϕ) distributions for each residue type. Four classes of mainchain conformation have been distinguished for each residue type. The utility of the two distributions in preparing starting structures for energy minimization using the Empirical Conformational Energy Program for Peptides (ECEPP/3) force field has been tested in two ways. First, the backbone with the correct mainchain classes was used to pick the most probable starting sidechain conformations, once each, according to the trivariate and bivariate distributions. The trivariate distribution performed better than the bivariate

distribution. Second, the sidechains were modelled on the backbone whose mainchain classes were predicted with only about 80% accuracy. In this case, it was better to use the most probable rotamer independent of (ψ,ϕ) than to use the most probable rotamer in the trivariate distribution. A probable reason for this is that the errors in the backbone prediction are larger than the sensitivity of the correlations between the mainchain and sidechain conformations. The conclusion is that the trivariate distribution is useful in sidechain modelling on a fixed or an almost fixed backbone only when a sufficiently accurate backbone is available.

Shenkin *et al.* [41••] describe their sidechain-modelling method that builds sidechains onto a fixed backbone. It relies on the probabilities from a rotamer library [47] extended to all χ_1 angles and also avoids atom–atom overlaps. The search is performed by a simulated annealing Monte Carlo procedure. A low temperature sampling around the optimal model is used to estimate statistical entropy for each sidechain conformation. This entropy is observed to correlate with the prediction accuracy, which allows an assignment of a confidence level to the prediction of each sidechain conformation. No correlation between rotamer entropies and solvent accessibilities is observed, as noted previously [48]. Shenkin *et al.* [41••] suggest that this may reflect approximately equal flexibility of both the buried and exposed sidechains. The method predicts the correct rotamer for 57% of the sidechains in a sample of 49 proteins; 74% of the χ_1 angles are found in the correct minimum. If half of the predictions with the lowest entropy are selected, the correct rotamer is obtained in 79% and the correct χ_1 in 84% of cases. All predictions are performed with the native backbone.

Vásquez [18•] reviews and comments on various approaches to sidechain modelling. He emphasizes the importance of two effects generally not taken into account: first, the coupling between mainchain and sidechains; and second, the continuous nature of the distributions of sidechain dihedral angles; for example, 5–30% of sidechains in crystal structures are significantly different from their rotamer conformations [49]. Both effects appear to be important when correlating packing energies and stability [39••]. The correct energetics may be obtained for the incorrect reasons; that is, the sidechains adopt distorted conformations to compensate for the rigidity of the backbone. Correspondingly, the backbone shifts may hinder the use of these methods when the template structures are related by less than 50% sequence identity [42••]. This is consistent with the X-ray structure of a variant of λ repressor that reveals that the protein accommodates the potentially disruptive residues with shifts in its α-helical arrangement and with only limited changes in sidechain orientations [50]. Some attempts to include backbone flexibility into sidechain modelling have been described [33,51] but are not yet generally applicable.

Chung and Subbiah [42••,43] give an elegant structural explanation for the rapid decrease in the conservation of sidechain packing as the sequence identity decreases below 30% [42••,43]. Although the fold is maintained, the pattern of sidechain interactions is generally lost in this range of sequence similarity [52]. Two sets of computations have been carried out for two sample protein sequences: the sidechain conformation has been predicted by maximizing packing both on the fixed native backbone and on a fixed backbone with approximately 2 Å rmsd from the native backbone; the 2 Å rmsd generally corresponds to about 25–30% sequence identity between two proteins. The sidechain predictions based on the two kinds of a backbone turn out to be unrelated. Thus, in as much as packing reflects the true laws determining sidechain conformation, a backbone with less than 30% sequence identity to the sequence being modelled is no longer sufficiently restraining to result in the correct packing of the buried sidechains.

Modelling of whole structures

The three different approaches to comparative modelling of the whole fold — modelling using assembly of rigid bodies, using segment matching, and using satisfaction of spatial restraints — are reviewed in [3] (see also the Introduction). All these approaches rely on an alignment between the target sequence and at least one template structure. In this review, we will not discuss the identification of the templates and their alignment with the target sequence [3,53,54], nor the early 1995 papers on comparative modelling [55•,56•] that are not reviewed in [3]. Some available software packages for comparative modelling of whole proteins are listed in Table 1.

Because the modelling 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-modelling techniques. For example, restraints could be provided by rules for secondary-structure packing [57], analyses of hydrophobicity [58••] and correlated mutations [59], empirical potentials of mean force [60], NMR experiments [61], cross-linking experiments [62], image reconstruction in electron microscopy [63], site-directed mutagenesis [64], 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.

Aszódi and Taylor [58••] describe a comparative-modelling method based on satisfaction of spatial constraints using distance geometry. The input is an alignment of the target sequence with the template structures, the output is a simplified model of the protein consisting of Cα atoms and pseudoatoms corresponding to a centroid of each sidechain. The procedure is implemented in program DRAGON, which is related to other methods based on distance geometry [14,15]. The procedure

Table 1

Available software packages for comparative modelling of whole proteins.

Program	Availability	World Wide Web address	Method*	Reference
COMPOSER	Public	http://felix.bioc.cam.ac.uk/soft-base/html	1	[100]
CONGEN	Public	email: bruc@dino.squibb.com	1	[101]
DRAGON	Public	http://www.nimr.mrc.ac.uk/~mathbio/a-aszodi/dragon.html	3	[58••]
MODELLER	Public	http://guitar.rockefeller.edu/modeller/modeller.html	3	[16]
NAOMI	Public	http://www.ocms.ox.ac.uk./~smb/Software/N_details/naomi.html	3	[102]
WHAT IF	Public	http://www.sander/.embl.heidelberg.de/vriend/	1	[103]
InsightII	Commercial	http://www.msi.com/	1, 3	(a)
LOOK	Commercial	http://www.mag.com/	2	[13]
QUANTA	Commercial	http://www.msi.com/	1, 3	(a)
SYBYL	Commercial	http://www.tripos.com/	1, 3	(b)
SWISS-MOD	Public server	http://www-isrec.unil.ch/SWISS-MODEL.html	1	[104]

*Method key (see also the Introduction): 1, comparative modelling by assembly of rigid bodies; 2, comparative modelling by segment matching; 3, comparative modelling by satisfaction of spatial restraints. SYBYL includes COMPOSER; QUANTA and InsightII include MODELLER, as well as in-house algorithms for modelling by rigid body assembly; InsightII also includes CONSENSUS [105]. SWISS-MOD is an Internet server for comparative modelling that takes either the target sequence or its alignment with known structures as input. Many additional programs specialize in modelling of sidechains or loops only. (a) Molecular Simulations Inc, San Diego. (b) Tripos, St Louis.

consists of gradually projecting a model of a protein from a high dimensional Euclidean space into the 3D dimensional space, subject to a large number of distance constraints. The distance constraints include homology-derived upper and lower distance bounds that are obtained from the alignment. These bounds are supplemented by stereochemical and more general distance restraints inferred from the conserved hydrophobicity patterns in the alignment. The method is rapid and might therefore be useful for generating preliminary models for the evaluation of alignments and for detailed refinement.

Accuracy of comparative models

Recently, protein modellers have been challenged to model sequences with unknown 3D structure again and to submit their models to the second 'Meeting on Critical Assessment of Techniques for Protein Structure Prediction' (CASP) in Asilomar [65]. At the same time, the 3D structures of the prediction targets were being determined by X-ray crystallography or NMR methods. The structures only became available after the models were calculated. Thus, it was possible to test the modelling methods objectively.

The best comparative techniques have been found to generally produce models with good stereochemistry and an overall structural accuracy that is slightly higher than the similarity between the template and the actual target structures when the modelling alignment is correct. Two modest improvements relative to the results from the first CASP meeting in 1994 [66] are apparent: better alignments resulting from more careful manual editing; and better techniques for modelling insertions shorter than about eight residues.

The errors in comparative models can be divided into four categories [67]: sidechain-packing errors; distortions and rigid-body changes in regions that are aligned correctly (e.g. loops, helices); distortions and rigid-body changes in insertions (e.g. loops); and distortions in incorrectly aligned regions (e.g. loops and longer segments that have low sequence identity to the templates).

The consequence of these errors is that the comparative method can result in models with a mainchain rms error as low as 1 Å for 90% of the mainchain residues, if a sequence is at least 40% identical to one or more of the templates [67]. 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 sidechains. 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 mainchain rms error rises to ~1.5 Å for about 80% of residues. The rest of the residues are modelled with large errors because the methods generally cannot model structural distortions and rigid-body shifts and cannot recover from misalignments (see Conclusions). Currently, insertions longer than about eight residues cannot usually be modelled accurately, whereas shorter loops can frequently be modelled successfully [17]. Model evaluation methods are frequently successful in identifying the inaccurately modelled regions of a protein [68].

When there are alignment errors in the template–target alignment used for modelling, 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 modelling 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 (R Sánchez, A Šali, unpublished data). 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 modelling applications.

To put the errors into perspective, we will list the differences among experimentally determined structures of the same protein. The 1 Å accuracy of mainchain atom positions corresponds to X-ray structures defined at a low resolution of about 2.5 Å and with an R-factor of about 25% [69], as well as to medium resolution NMR structures determined from ten interproton distance restraints per residue [70,71]. Similarly, differences between the highly refined X-ray and NMR structures of the same protein also tend to be about 1 Å [70]. Changes in the environment (e.g. crystal packing, solvent, ligands) can also have a significant effect on the structure [72]. Overall, comparative modelling based on templates with more than 40% identity is almost as good, simply because the homologues 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 modelling, however, is that some regions, mainly loops and sidechains, have larger errors.

Comparative modelling and genome databases

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. For many protein sequences, however, comparative modelling can provide a useful 3D model. In fact, about one third of the 198,449 known protein sequences [73] are related to at least one of the 4,861 known protein structures (Brookhaven Protein Data Bank [74,75]) [4•,76,77,78•]. Thus, the number of sequences that can be modelled relatively accurately at this moment is an order of magnitude larger than the number of experimentally determined protein structures. Furthermore, the usefulness of comparative modelling 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 [76,77,78•]. 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 six years (see Fig. 4c in [78•]), thus allowing comparative modelling to be applicable to most of the protein sequences at that time. This means that when the human genome project is finished it will be possible to use comparative modelling to obtain approximate 3D models for most of the proteins in the human genome.

Comparative modelling, 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 [79], substrate specificity [80•], and drug design [81]. It can also provide starting models in X-ray crystallography [82] and NMR spectroscopy [61]. Possible uses of comparative modelling are illustrated by a number of recent applications [62,80•,83•,84•,85,86,87•,88]). An exhaustive survey of many comparative-modelling studies is given in [1].

Conclusions

Future improvements of comparative modelling should aim to model proteins with lower similarities to known structures (e.g. <30% sequence identity), to increase the accuracy of the models, and to make modelling fully automated. The improvements will probably include the simultaneous optimization of sidechain and backbone conformations in sidechain modelling, the simultaneous optimization of a loop and its environment in loop modelling, and the 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 [16], could be one way of achieving this goal [89••,90••]. 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 mainchain rmsd from the X-ray structure smaller than about 1 Å, even when the starting conformation is the X-ray structure itself. For example, MD simulations in solvent generally have a mainchain rmsd of more than 1 Å and the most detailed lattice folding simulations result in models with an rms error larger than 2 Å [91]. As most of the mainchain atoms in two homologues with at least 40% sequence identity usually superimpose with an rmsd of about 1 Å, it is currently better to aim to reproduce the template structures as closely as possible, rather than to venture away from the templates in the search for a better model.

The major factor that limits the use of comparative modelling 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 decreases in sequence similarity [92]; an additional complication is that even structure–structure comparisons may not result in a unique alignment for proteins with less than about 25% identity [93•,94•]. This limits the usefulness of comparative modelling because no current modelling technique can recover from an incorrect input alignment. Profile matching [95] and threading methods [96–98] appear to be a natural solution to the alignment problem in comparative modelling. Whereas these techniques are successful in identifying related folds, however, they currently appear to be somewhat less successful in generating correct alignments. To reduce the errors in the model that stemm from the alignment errors, iterative changes in the alignment during the calculation of the model are needed. A case in point is provided by the generation and analysis of multiple models, based on different templates, for the EF-hand calcium-binding proteins [99•].

Although comparative modelling needs significant improvement, it is already a mature technique that can be used to address many practical problems. With the increase in the number of protein sequences discovered and in the fraction of all folds that are known, comparative modelling will be an increasingly important tool for biologists who seek to understand and control normal and disease-related processes in living organisms.

Acknowledgements

Andrej Šali is a Sinsheimer Scholar. Roberto Sánchez is a Howard Hughes Medical Institute predoctoral fellow. Support by National Institutes of Health grant GM 54762 is also acknowledged.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Johnson MS, Srinivasan N, Sowdhamini R, Blundell BL: Knowledge-based protein modelling. *CRC Crit Rev Biochem Mol Biol* 1994, 29:1–68.

2. Bajorath J, Stenkamp R, Aruffo A: Knowledge-based model building of proteins: concepts and examples. *Protein Sci* 1994, 2:1798–1810.

3. Šali A: Modelling mutations and homologous proteins. *Curr Opin Biotechnol* 1995, 6:437–451.

4. • Rost B, Sander C: Bridging the protein sequence-structure gap by structure predictions. *Annu Rev Biophys Biomol Struct* 1996, 25:113–136.
A review of protein structure prediction methods, including secondary-structure prediction, contact prediction, comparative modelling and fold recognition.

5. Lesk AM, Chothia CH: The response of protein structures to amino-acid sequence changes. *Phil Trans R Soc London Ser B* 1986, 317:345–356.

6. Hubbard TJP, Blundell TL: Comparison of solvent inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Eng* 1987, 1:159–171.

7. Browne WJ, North ACT, Phillips DC, Brew K, Vanaman TC, Hill RC: A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. *J Mol Biol* 1969, 42:65–86.

8. Greer J: Comparative model-building of the mammalian serine proteases. *J Mol Biol* 1981, 153:1027–1042.

9. Blundell TL, Sibanda BL, Sternberg MJE, Thornton JM: Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* 1987, 326:347–352.

10. Jones TH, Thirup S: Using known substructures in protein model building and crystallography. *EMBO J* 1986, 5:819–822.

11. Unger R, Harel D, Wherland W, Sussman JL: A 3-D building blocks approach to analyzing and predicting structure of proteins. *Proteins* 1989, 5:355–373.

12. Claessens M, Cutsem EV, Lasters I, Wodak S: Modelling the polypeptide backbone with 'spare parts' from known protein structures. *Protein Eng* 1989, 4:335–345.

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

14. Havel TF, Snow ME: A new method for building protein conformations from sequence alignments with homologues of known structure. *J Mol Biol* 1991, 217:1–7.

15. Srinivasan S, March CJ, Sudarsanam S: An automated method for modeling proteins on known templates using distance geometry. *Protein Sci* 1993, 2:227–289.

16. Šali A, Blundell TL: Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 1993, 234:779–815.

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

18. • Vásquez M: Modeling side-chain conformation. *Curr Opin Struct Biol* 1996, 6:217–221.
The main approximations in sidechain modelling with emphasis on the rigid-backbone approximation are reviewed.

19. Moult J, James MNG: An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins* 1986, 1:146–163.

20. Bruccoleri RE, Karplus M: Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* 1987, 26:137–168.

21. Fine RM, Wang H, Shenkin PS, Yarmush DL, Levinthal C: Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of MCP603 from many randomly generated loop conformations. *Proteins* 1986, 1:342–362.

22. Martin ACR, Cheetham JC, Rees AR: Modeling antibody hypervariable loops: a combined algorithm. *Proc Natl Acad Sci USA* 1989, 86:9268–9272.

23. Chothia C, Lesk AM, Levitt M, Amit AG, Mariuzza RA, Phillips SEV, Poljak RJ: The predicted structure of immunoglobulin d1.3 and its comparison with the crystal structure. *Science* 1986, 233:755–758.

24. • Tenette C, Ducancel F, Smith JC: Structural model of the anti-snake-toxin antibody, Mα2,3. *Proteins* 1996, 26:9–31.
A combination of three existing methods to model antibody CDR loops is described.

25. Reczko M, Martin ACR, Bohr H, Suhai S: Prediction of hypervariable CDR-H3 loop structures in antibodies. *Protein Eng* 1996, 8:389–395.

26. •• Zheng Q, Kyle DJ: Accuracy and reliability of the scaling-relaxation method for loop closure: an evaluation based

on extensive and multiple copy conformational samplings. *Proteins* 1996, 24:209–217.
An extension and evaluation of the previously published loop-modelling method is described [27].

27. Zheng Q, Rosenfeld RJ, Vajda S, DeLisi C: Determining protein loop conformation using scaling-relaxation techniques. *Protein Sci* 1993, 2:1242–1248.

28. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M: CHARMm: a program for macromolecular energy minimization and dynamics calculations. *J Comp Chem* 1983, 4:187–217.

29. •• Donate LE, Rufino SD, Canard LHJ, Blundell TL: Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction. *Protein Sci* 1996, 5:2600–2616.
A classification of loops by length, sequence, mainchain conformation and relative disposition of the flanking secondary-structure segments is described.

30. • Kwasigroch J-M, Chomilier J, Mornon J-P: A global taxonomy of loops in globular proteins. *J Mol Biol* 1996, 259:855–872.
A database of loops that are clustered into families on the basis of their length and conformation is described.

31. Nayeem A, Scheraga HA: A statistical analysis of side-chain conformations in proteins: comparison between ECEPP predictions. *J Protein Chem* 1994, 13:283–296.

32. Goldstein RF: Efficient rotamer elimination applied to protein side chains and related spin glasses. *Biophys J* 1994, 66:1335–1340.

33. Harbury PB, Tidor B, Kim PS: Repacking proteins cores with backbone freedom: structure prediction for coiled coils. *Proc Natl Acad Sci USA* 1995, 92:8408–8412.

34. Lasters I, De Maeyer M, Desmet J: Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side chains. *Protein Eng* 1995, 8:815–822.

35. Vásquez M: An evaluation of discrete and continuum search techniques for conformational analysis of side chains in proteins. *Biopolymers* 1995, 36:53–70.

36. Keller DA, Shibata M, Marcus E, Ornstein RL, Rein R: Finding the global minimum: a fuzzy end elimination implementation. *Protein Eng* 1995, 8:893–904.

37. Hwang JK, Liao WF: Side-chain prediction by neural networks and simulated annealing optimization. *Protein Eng* 1995, 8:893–904.

38. •• Bower MJ, Cohen FE, Dunbrack RL: Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J Mol Biol* 1997, in press.
A backbone-dependent rotamer library is tested in 'real world' comparative modelling cases using templates in the range from 30–90% sequence identity. A high prediction accuracy is obtained.

39. •• Lee C: Testing homology modeling on mutant proteins: predicting structural and thermodynamic effects in the Ala98→Val mutants of T4 lysozyme. *Fold Des* 1996, 1:1–12.
Two variants of a packing method for modelling sidechains on a fixed backbone are tested on single mutants of T4 lysozyme. The impact of the backbone accuracy on structural and thermodynamic predictions is assessed. The method that allows finer steps in χ_1 torsion rotations is shown to be less sensitive to backbone errors than the rotamer-based model.

40. •• Cheng B, Nayeem A, Scheraga HA: From secondary structure to three-dimensional structure: improved dihedral angle probability distribution function for use with energy searches for native structures of polypeptides and proteins. *J Comp Chem* 1996, 17:1453–1480.
Trivariate (ψ,ϕ,χ_1) probability distributions are derived for each residue type from a database of known protein structures. Their performance in mainchain/sidechain conformational searches is compared with those of some alternative probability distributions.

41. •• Shenkin SP, Farid H, Fetrow JS: Prediction and evaluation of side-chain conforma-tions for protein backbone structures. *Proteins* 1996, 26:323–352.
A rapid algorithm for sidechain rotamer prediction on a fixed backbone is described. It is based on a rotamer library and Monte Carlo search. A method for estimating prediction accuracy is proposed.

42. •• Chung SY, Subbiah S: A structural explanation for the twilight zone of protein sequence homology. *Structure* 1996, 4:1123–1127.
A possible structural explanation is given for the abrupt and large changes in the sidechain-packing patterns when mainchain differences become larger than approximately 2 Å.

43. Chung SY, Subbiah S: The use of side-chain packing methods in modeling bacteriophage repressor and cro proteins. *Protein Sci* 1995, 4:2300–2309.

44. The backbone-dependent rotamer library on World Wide Web URL: http://www.cmpharm.ucsf.edu/~dunbrack

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

46. Koehl P, Delarue M: Mean-field minimization methods for biological macromolecules. *Curr Opin Struct Biol* 1996, 6:222–226.

47. Ponder JW, Richards FM: Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol* 1987, 193:775–791.

48. Koehl P, Delarue M: Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J Mol Biol* 1994, 239:249–275.

49. Schrauber H, Eisenhaber F, Argos P: Rotamers: to be or not to be? An analysis of amino acid side-chain conformations in globular proteins. *J Mol Biol* 1993, 230:592–612.

50. Lim WA, Hodel A, Sauer RT, Richards FM: The crystal structure of a mutant protein with altered but improved hydrophobic core packing. *Proc Natl Acad Sci USA* 1994, 91:423–427.

51. Koehl P, Delarue M: A self consistent mean field approach to simultaneous gap closure and side-chain positioning in protein homology modelling. *Nat Struct Biol* 1995, 2:163–170.

52. Russell RB, Barton GJ: Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *J Mol Biol* 1994, 244:332–350.

53. Holm L, Sander C: Searching protein structure databases has come of age. *Proteins* 1994, 19:165–173.

54. Barton GJ: Protein sequence alignment and database scanning. In *Protein Structure Prediction: A Practical Approach.* Edited by Sternberg MJE. Oxford: IRL Press; 1997:31–63.

55. • Lipke PN, Chen M-H, De Nobel H, Kurjan J, Kahn PC: Homology modeling of an immunoglobulin-like domain in the ***Saccharomyces cerevisiae*** adhesion protein α-agglutinin. *Protein Sci* 1995, 4:2168–2178.
An example of a difficult comparative modelling case, which is based on a low similarity between the target and template proteins. The secondary-structure predictions are used systematically to improve the input alignment for the modelling procedure.

56. • Mandal C, Kingery BD, Anchin JM, Subramaniam S, Linthicum DS: ABGEN: a knowledge-based automated approach for antibody structure modeling. *Nat Biotechnol* 1996, 14:323–328.
A completely automated antibody-modelling system that uses the existing sequence and structural databases of known antibodies is described.

57. Taylor WR: Protein fold-refinement: building models from idealized folds using motif constraints and multiple sequence data. *Protein Eng* 1993, 6:593–604.

58. •• Aszódi A, Taylor WR: Homology modelling by distance geometry. *Fold Des* 1996, 1:325–334.
A description of a comparative-modelling method that is based on distance geometry is presented. The method is useful for the quick generation of a large number of low resolution models. The models consist of residues described by a Cα atom and a sidechain pseudoatom.

59. Gobel U, Sander C, Schneider R, Valencia A: Correlated mutations and residue contacts in proteins. *Proteins* 1994, 18:309–317.

60. Sippl MJ: 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* 1990, 213:859–883.

61. Sutcliffe MJ, Dobson CM, Oswald RE: Solution structure of neuronal bungaro-toxin determined by two-dimensional NMR spectroscopy: calculation of tertiary structure using systematic homologous model building, dynamical simulated annealing, and restrained molecular dynamics. *Biochemistry* 1992, 31:2962–2970.

62. Rossi V, Gaboriaud C, Lacroix M, Ulrich J, Fontecilla-Camps JC, Gagnon J, Arlaud GJ: Structure of the catalytic region of human complement protease c1s: study by chemical cross-linking and three-dimensional homology modelling. *Biochemistry* 1995, 34:7311–7321.

63. Neil KJ: Structure of recombinant rat UBF by electron image analysis and homology modelling. *Nucleic Acids Res* 1995, 24:1472–1480.

64. Boissel JP, Lee WR, Presnell SR, Cohen FE, Bunn HF: Erythropoietin structure–function relationships. Mutant proteins that test a model of tertiary structure. *J Biol Chem* 1993, 268:15983–15993.

65. Second meeting on the critical assessment of techniques for protein structure prediction on World Wide Web URL: http://iris4.carb.nist.gov/casp2/

66. Mosimann S, Meleshko R, James MNG: A critical assessment of comparative molecular modeling of tertiary structures of proteins. *Proteins* 1995, 23:301–317.

67. Šali A, Potterton L, Yuan F, Van Vlijmen H, Karplus M: Evaluation of comparative protein modeling by MODELLER. *Proteins* 1995, 23:318–326.

68. Sippl MJ: Recognition of errors in three-dimensional structures of proteins. *Proteins* 1993, 17:355–362.

69. Ohlendorf DH: Accuracy of refined protein structures. II. Comparison of four independently refined models of human interleukin 1β. *Acta Crystallogr D* 1994, 50:808–812.

70. Clore GM, Robien MA, Gronenborn AM: Exploring the limits of precision and accuracy of protein structures determined by nuclear magnetic resonance spectroscopy. *J Mol Biol* 1993, 231:82–102.

71. Zhao D, Jardetzky O: An assessment of the precision and accuracy of protein structures determined by NMR. *J Mol Biol* 1994, 239:601–607.

72. Faber HR, Matthews BW: A mutant T4 lysozyme displays five different crystal conformations. *Nature* 1990, 348:263–266.

73. Database growth on World Wide Web URL: http://www.dna.affrc.go.jp/htdocs/growth/

74. Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J: Protein Data Bank. In *Crystallographic Databases Information, Content, Software Systems, Scientific Applications.* Edited by Allen FH, Bergerhoff G, Sievers R. Bonn/Cambridge/Chester: Data Commission of the International Union of Crystallography; 1987:107–132.

75. Protein data bank on World Wide Web URL: http://www.pdb.bnl.gov/

76. Chothia C: One thousand families for the molecular biologist. *Nature* 1992, 360:543–544.

77. Orengo CA, Jones DT, Thornton JM: Protein superfamilies and domain superfolds. *Nature* 1994, 372:631–634.

78. • Holm L, Sander C: Mapping the protein universe. *Science* 1996, 273:595–602.
A review of the use of protein-structure comparisons in protein classification and in function identification.

79. Matsumoto R, Šali A, Ghildyal N, Karplus M, Stevens RL: 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* 1995, 270:19524–19531.

80. • Xu LZ, Sánchez R, Šali A, Heintz N: Ligand specificity of brain lipid binding protein. *J Biol Chem* 1996, 271:24711–24719.
A comparative model of brain lipid-binding protein is used to suggest mutations for exploring ligand specificity. The mutants are expressed and validated experimentally.

81. Ring CS, Sun E, McKerrow JH, Lee GK, Rosenthal PJ, Kuntz ID, Cohen FE: Structure-based inhibitor design by using protein models for the development of antiparasitic agents. *Proc Natl Acad Sci USA* 1993, 90:3583–3587.

82. Carson M, Bugg CE, Delucas L, Narayana S: Comparison of homology models with the experimental structure of a novel serine protease. *Acta Crystallogr D* 1994, 50:889–899.

83. • Modi S, Paine MJ, Sutcliffe MJ, Lian L-Y, Primrose WU, Wolf CR, Roberts GCK: A model for human cytochrome $p_4$50 2d6 based on homology modeling and NMR studies of substrate binding. *Biochemistry* 1996, 35:4540–4550.
An example of how NMR-derived restraints can be combined with comparative modelling to improve the resulting protein model.

84. • Chen X, Whitmire D, Bowen JP: Xylanase homology modelling using the inverse protein folding approach. *Protein Sci* 1996, 5:705–708.
An example of the use of threading to obtain a remotely related structure for comparative modelling.

85. Adzhubei AA, Laughton CA, Neidle S: An approach to protein homology modelling based on an ensemble of NMR structures; application to the Sox-5 HMG-box protein. *Protein Eng* 1995, 8:615–625.

86. Loew GH, Du P, Smith AT: Homology modelling of horseradish peroxidase coupled to two-dimensional NMR spectral assignments. *Biochem Soc Trans* 1995, 23:250–256.

87. • Ott K-H, Kwagh J-G, Stockton GW, Sidorov V, Kakefuda G: Rational molecular design and genetic engineering of herbicide resistant crops by structure modeling and site-directed mutagenesis of acetohydroxyacid synthase. *J Mol Biol* 1996, 263:359–368.
A successful application of comparative modelling to solve a real life problem.

88. • Bajorath J, Sheriff S: Comparison of an antibody model with an X-ray structure; the variable fragment of BR96. *Proteins* 1996, 24:152–157.
The BR96 antibody model is compared with two X-ray structures determined after the model was built. Most of the residues important for interaction with an antigen were identified correctly.

89. •• Sippl MJ, Ortner M, Jaritz M, Lackner P, Flöckner H: Helmholtz free energies of atom pair interactions in proteins. *Fold Des* 1996, 1:275–288.

Statistical potentials for atom–atom distances are described. Potentials like this one will probably be useful in improving the accuracy of comparative modelling.

90. •• DeBolt SE, Skolnick J: Evaluation of atomic level mean force potentials via inverse folding and inverse refinement of protein structures: atomic burial position and pairwise non-bonded interactions. *Protein Eng* 1996, 9:937–955.
Statistical potentials that could in principle be used to improve the accuracy of comparative models are described.

91. Kolinski A, Skolnick J: Monte Carlo simulations of protein folding. II Application to protein A, ROP, and crambin. *Proteins* 1994, 18:353–366.

92. Johnson MS, Overington JP: A structural basis for sequence comparisons: an evaluation of scoring methodologies. *J Mol Biol* 1993, 233:716–738.

93. • Zu-Kang F, Sippl MJ: Optimum superposition of protein structures: ambiguities and implications. *Fold Des* 1996, 1:123–132.
See annotation [94•].

94. • Godzik A: The structural alignment between two proteins: is there a unique answer? *Protein Sci* 1996, 5:1325–1338.
These two papers [93•,94•] show that a unique structure–structure alignment is not always guaranteed when the compared structures are sufficiently different.

95. Bowie JU, Lütthy R, Eisenberg D: A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 1991, 253:164–170.

96. Finkelstein AV, Reva BA: A search for the most stable folds of protein chains. *Nature* 1991, 351:497–499.

97. Jones DT, Taylor WR, Thornton JM: A new approach to protein fold recognition. *Nature* 1992, 358:86–89.

98. Godzik A, Kolinski A, Skolnick J: Topology fingerprint approach to the inverse protein folding problem. *J Mol Biol* 1992, 227:227–238.

99. • Pawlowski K, Bierzynski A, Godzik A: Structural diversity in a family of homologous proteins. *J Mol Biol* 1996, 258:349–366.
Comparative models for the EF-hand calcium-binding proteins are calculated from a number of different templates. The models are compared with the actual structures and evaluated with respect to several energetic criteria. Putative energy factors responsible for the selection of the particular quaternary structure among several alternatives are identified.

100. Sutcliffe MJ, Haneef I, Carney C, Blundell TL: Knowledge based modelling of homologous proteins. Part I: three dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng* 1987, 1:377–384.

101. Bruccoleri RE: Application of systematic conformational search to protein modeling. *Molecular Simulation* 1993, 10:151–174.

102. Brocklehurst SM, Perham RN: 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 methods for the prediction of protein tertiary structure. *Protein Sci* 1993, 2:626–639.

103. Vriend G: WHAT IF: a molecular modeling and drug design program. *J Mol Graph* 1990, 8:52–56.

104. Peitsch MC, Jongeneel CV: A 3-D model for the CD40 ligand predicts that it is a compact trimer similar to the tumor necrosis factors. *Int Immunol* 1993, 5:233–238.

105. Havel TF: Predicting the structure of the flavodoxin from ***Escherichia coli*** by homology modeling, distance geometry and molecular dynamics. *Mol Simulation* 1993, 10:175–210.